# Supplementary material for "Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding"

In Appendix A, we present the proof of Theorem 1 and important intermediary results for establishing Theorem 1. In Appendix B, we present the proof of Proposition 1, which relies on a finite-sample analysis of the factor model. Additional Proofs are presented in Appendix C.

## APPENDIX A: INTERMEDIARY RESULTS AND PROOF OF THEOREM 1

In the following, we list three intermediary results in Sections A.1 to A.3 as the key components of proving our main result Theorem 1 and then provide the proof of Theorem 1 in Section A.4. We verify the condition (A2) in Section A.5. All our theoretical derivations are done for the Hidden Confounding Model (2), but they additionally hold more generally for the perturbed linear model (3).

**A.1. Valid spectral transformations.** The first intermediary result is on the properties of the spectral transformation we use. We will show that the limiting distribution in Theorem 1 holds generally for the estimator (10) using any spectral transformations $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ that satisfy the following:

(P1) **Spectral Transformation Property.** $\mathcal{P}^{(j)} = U(X_{-j})S(X_{-j})U(X_{-j})^\intercal$ and $\mathcal{Q} = U(X)S(X)U(X)^\intercal$ satisfy

$$\frac{1}{n}\|\mathcal{P}^{(j)}X_{-j}\|_2^2 \lesssim \max\left\{1,\frac{p}{n}\right\} \quad \text{and} \quad \frac{1}{n}\|\mathcal{Q}X\|_2^2 \lesssim \max\left\{1,\frac{p}{n}\right\} \tag{30}$$

$$\mathrm{Tr}[(\mathcal{P}^{(j)})^4] = \sum_{l=1}^{n}[S_{l,l}(X_{-j})]^4 \gtrsim m \quad \text{and} \quad \mathrm{Tr}(\mathcal{Q}^4) = \sum_{l=1}^{n}[S_{l,l}(X)]^4 \gtrsim m. \tag{31}$$

with $m = \min\{n, p-1\}$.

The first requirement means that $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ need to shrink the leading singular values of $X_{-j}$ and $X$ to a sufficiently small level, respectively. On the other hand, the second requirement says that the overall shrinkage of all singular values together is not too big.

For the proof of Theorem 1 and its intermediate results, we extensively use that our spectral transformations satisfy the property (P1). Therefore, we first need to show that the Trim transform $\mathcal{P}^{(j)}$ defined in (14) and $\mathcal{Q}$ defined in (15) satisfy the property (P1). Since $S_{l,l} = 1$ for $l > \lfloor \rho m \rfloor$, we have that at least $\lfloor (1-\rho)m \rfloor$ diagonal elements of $S$ are equal to 1, which immediately gives us (31) for $\mathcal{Q}$ whenever $\rho < 1$. Similarly, (31) for $\mathcal{P}^{(j)}$ holds for any $\rho_j \in (0,1)$. However, in order to show the condition (30), we need to better understand the behaviour of the singular values of the random matrix $X$.

PROPOSITION 3. *Suppose $E_{i,\cdot} \in \mathbb{R}^p$ is a sub-Gaussian random vector and $\lambda_{\max}(\Sigma_E) \leq C$, for some positive constant $C > 0$, then with probability larger than $1 - \exp(-cn)$,*

$$\lambda_{q+1}\left(\tfrac{1}{n}X^\intercal X\right) \lesssim \max\{1, p/n\},$$

*for some positive constant $c > 0$.*

The above proposition is proved in the Section C.2 by applying the Weyl's inequality. This now allows us to conclude that the Trim transform satisfies the property (P1):

COROLLARY 2. *Let $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ be the spectral transformation matrices obtained by applying the Trim transformation (14) and (15), respectively. Suppose that the conditions of Proposition 3 hold and that $\min\{\rho, \rho_j\} \geq (q+1)/\min\{n, p-1\}$ and $\max\{\rho, \rho_j\} < 1$. Then the Trim transformations $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ satisfy* (P1).

**A.2. Approximate sparsity and perturbation size.** The essential step of bias correction is to decouple the correlation between the variable of interest $X_{1,j}$ and other covariates $X_{1,-j} \in \mathbb{R}^{p-1}$. In order to get an informative projection direction $\mathcal{P}^{(j)} Z_j$, one needs to estimate the best linear approximation vector $\gamma = [\mathbb{E}(X_{1,-j} X_{1,-j}^{\intercal})]^{-1} \mathbb{E}(X_{1,-j} X_{1,j}) \in \mathbb{R}^{p-1}$ well. Recall that the results for the standard Debiased Lasso [56] are based on the fact that the sparsity of the precision matrix $\Sigma_X^{-1}$ gives sparsity of $\gamma$, thus justifying the estimation accuracy of the Lasso regression of $X_{1,j}$ on $X_{1,-j}$. However, even though the assumption (A1) ensures the sparsity of the precision matrix of the unconfounded part $E$, $\gamma$ will not be sparse, since the confounding variables $H$ introduce additional correlations between the covariates $X$.

Recall the definitions

$$\eta_{i,j} = X_{i,j} - X_{i,-j}^{\intercal} \gamma \quad \text{and} \quad \nu_{i,j} = E_{i,j} - E_{i,-j}^{\intercal} \gamma^E,$$

where $\gamma^E = [\mathbb{E}(E_{1,-j} E_{1,-j}^{\intercal})]^{-1} \mathbb{E}(E_{1,-j} E_{1,j})$.

The following Lemma 1 shows that in the presence of confounding variables, the vector $\gamma$ can be decomposed into a main sparse component $\gamma^E$ and an additional small perturbation vector $\gamma^A$. The proof of the following Lemma is presented in Section C.3.

LEMMA 1. *Suppose that the conditions* (A1) *and* (A2) *hold, then the vector* $\gamma = [\mathbb{E}(X_{1,-j} X_{1,-j}^{\intercal})]^{-1} \mathbb{E}(X_{1,-j} X_{1,j})$ *defined as the minimizer of* $\mathbb{E}(X_{1,j} - X_{1,-j}^{\intercal} \gamma')^2$*, can be decomposed as* $\gamma = \gamma^E + \gamma^A$*, where* $\gamma^E = [\mathbb{E}(E_{1,-j} E_{1,-j}^{\intercal})]^{-1} \mathbb{E} E_{1,j} E_{1,-j}$ *is a sparse vector with at most* $s$ *non-zero components and the approximation error* $\gamma^A$ *satisfies*

$$\|\gamma^A\|_2 \leq \max_{1 \leq l \leq q} \frac{C_0 |\lambda_l(\Psi_{-j})|}{c_0 \lambda_l^2(\Psi_{-j}) + 1} \|\Psi_j + \Psi_{-j} \gamma^E\|_2 \lesssim \frac{\sqrt{q} (\log p)^{1/4}}{\lambda_q(\Psi_{-j})}. \tag{32}$$

*Furthermore, the difference* $\delta_{i,j} = \eta_{i,j} - \nu_{i,j}$ *satisfies*

$$\mathrm{Var}(\delta_{i,j}) \lesssim \frac{\|\Psi_j - \Psi_{-j} \gamma^E\|_2^2}{1 + \lambda_q^2(\Psi_{-j})} \lesssim \frac{q (\log p)^{1/2}}{1 + \lambda_q^2(\Psi_{-j})}. \tag{33}$$

The main component $\gamma^E$ is fully determined by the covariance structure of $E_{i,\cdot}$. From the block matrix inverse formula, we get that $\gamma^E$ is proportional to $(\Omega_E)_{j,-j} \in \mathbb{R}^{p-1}$ and therefore sparse with at most $s$ non-zero components. Since the additional component $\gamma^A$ converges to zero as in (32), the regression vector $\gamma$ is approximately sparse.

In a similar fashion, we will show that the perturbation $b$ in (3), which is induced by the confounding variables, is of a small order of magnitude as well.

LEMMA 2. *Suppose that the conditions* (A1) *and* (A2) *hold, then*

$$|b_j| \lesssim \frac{q (\log p)^{1/2}}{1 + \lambda_q^2(\Psi)}, \quad \|b\|_2 \lesssim \frac{\sqrt{q} (\log p)^{1/4}}{\lambda_q(\Psi)}, \tag{34}$$

*and*

$$\left|\sigma_\epsilon^2 - \sigma_e^2\right| = \left|\phi^{\intercal} \left(\mathrm{I}_q - \Psi \Sigma_X^{-1} \Psi^{\intercal}\right) \phi\right| \lesssim \frac{q (\log p)^{1/2}}{1 + \lambda_q^2(\Psi)}. \tag{35}$$

The above lemma also shows that the variance of the error $\epsilon_i$ in (3) is close to that of the random error $e_i$. The proof of the above lemma is presented in Section C.4.

**A.3. Error rates of $\widehat{\beta}^{init}$ and $\widehat{\gamma}$.** In order to show the asymptotic normality of the proposed Doubly Debiased Lasso estimator (10), we need that the estimators $\widehat{\beta}^{init}$ and $\widehat{\gamma}$ estimate the target values $\beta$ and $\gamma$ well. In the following proposition, we show that the estimator $\widehat{\gamma}$ described in (9) accurately estimates $\gamma$ with a high probability. The proof of Proposition 4 is presented in Section C.5.

PROPOSITION 4. *Suppose that the conditions* (A1) − (A4) *hold. If the spectral transformation* $\mathcal{P}^{(j)}$ *satisfies* (P1) *and the tuning parameter* $\lambda_j$ *in* (9) *is chosen as* $\lambda_j \geq A\sigma_j\sqrt{\frac{\log p}{n}} + \sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}}$, *for some positive constant* $A > 0$, *then with probability larger than* $1 - e\cdot p^{1-c(A/C_1)^2} - \exp(-cn) - (\log p)^{-1/2}$ *for some positive constant* $c > 0$, *the estimator* $\widehat{\gamma}$ *proposed in* (9) *satisfies*

$$\|\widehat{\gamma}-\gamma^E\|_1 \lesssim \|W_{-j,-j}(\widehat{\gamma}-\gamma^E)\|_1 \lesssim \frac{M^2}{\tau_*}s\lambda_j + \frac{1}{\lambda_j}\frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2}{n}, \tag{36}$$

$$\|\widehat{\gamma}-\gamma^E\|_2 \lesssim \frac{M}{\tau_*}\sqrt{s}\lambda_j + \frac{1}{\lambda_j}\frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2}{n}, \tag{37}$$

$$\frac{1}{\sqrt{n}}\|\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E)\|_2 \lesssim \frac{M}{\tau_*}\sqrt{s}\lambda_j + \frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2}{\sqrt{n}}, \tag{38}$$

*where* $W\in\mathbb{R}^{p\times p}$ *as a diagonal matrix with diagonal entries as* $W_{l,l} = \|\mathcal{P}^{(j)}X_{\cdot,l}\|_2/\sqrt{n}$ *for* $1\leq l\leq p$, $\tau_* > 0$ *is the lower bound for the restricted eigenvalue defined in* (22) *and* $M$ *is the sub-Gaussian norm for components of* $X_{i,\cdot}$, *as defined in Assumption (A3).*

Throughout our analysis, we shall choose $\lambda_j$ as

$$\lambda_j \asymp A\sigma_j\sqrt{\frac{\log p}{n}} + \sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}}, \tag{39}$$

though Proposition 4 shows that the results also hold for a larger $\lambda_j$. Furthermore, we combine (30) and (32) and establish

$$\frac{1}{n}\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2 \lesssim \max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{\lambda_q^2(\Psi_{-j})}. \tag{40}$$

In addition, we show an analogous result that the initial spectral deconfounding estimator $\widehat{\beta}^{init}$ proposed in (16) accurately estimates $\beta$ with a high probability:

PROPOSITION 5. *Suppose that the conditions* (A1) − (A4) *hold. If the spectral transformation* $\mathcal{Q}$ *satisfies* (P1) *and the tuning parameter* $\lambda$ *in* (16) *is chosen as* $\lambda \geq A\sigma_e\sqrt{\frac{\log p}{n}} + \sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi)}}$, *for some positive constant* $A > 0$, *then with probability larger than* $1 - e\cdot p^{1-c(A/C_1)^2} - \exp(-cn) - (\log p)^{-1/2}$ *for some positive constant* $c > 0$, *the estimator* $\widehat{\beta}^{init}$ *proposed in* (16) *satisfies*

$$\|\widehat{\beta}^{init}-\beta\|_1 \lesssim \|\widetilde{W}(\widehat{\beta}^{init}-\beta)\|_1 \leq \frac{M^2}{\tau_*}k\lambda + \frac{1}{\lambda}\frac{\|\mathcal{Q}Xb\|_2^2}{n}, \tag{41}$$

$$\|\widehat{\beta}^{init}-\beta\|_2 \leq \frac{M}{\tau_*}\sqrt{k}\lambda + \frac{1}{\lambda}\frac{\|\mathcal{Q}Xb\|_2^2}{n}, \tag{42}$$

$$\frac{1}{\sqrt{n}}\|\mathcal{Q}X(\widehat{\beta}^{init}-\beta)\|_2 \leq \frac{M}{\tau_*}\sqrt{k}\lambda + \frac{\|\mathcal{Q}Xb\|_2}{\sqrt{n}}, \tag{43}$$

*where $\widetilde{W} \in \mathbb{R}^{p\times p}$ as a diagonal matrix with diagonal entries as $\widetilde{W}_{l,l} = \|\mathcal{Q}X_{\cdot,l}\|_2/\sqrt{n}$ for $1 \leq l \leq p$, $\tau_* > 0$ is the lower bound for the restricted eigenvalue defined in (21) and $M$ is the sub-Gaussian norm for components of $X_{i,\cdot}$, as defined in Assumption (A3)..*

This extends the results in [12], where only the rate of convergence of $\|\widehat{\beta}^{init}-\beta\|_1$ has been established, but not of $\|\widehat{\beta}^{init}-\beta\|_2$ and $\frac{1}{\sqrt{n}}\|\mathcal{Q}X(\widehat{\beta}^{init}-\beta)\|_2$ and furthermore, the assumption (A2) is weaker than the assumption $\lambda_q(\Psi) \gtrsim \sqrt{p}$ required in Theorem 1 of [12]. The proof of Proposition 5 is presented in Section C.6. We shall choose

$$\lambda \asymp A\sigma_e\sqrt{\frac{\log p}{n}} + \sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi)}},$$

though Proposition 5 shows that the results also hold for a larger $\lambda$. Furthermore, similar to (40), we combine (30) and (34) and establish

$$\frac{1}{n}\|\mathcal{Q}Xb\|_2^2 \lesssim \max\left\{1, \frac{p}{n}\right\} \cdot \frac{q\sqrt{\log p}}{\lambda_q^2(\Psi)}. \tag{44}$$

As a remark, if we further assume the error $\epsilon_i$ in the model (3) to be independent of $X_{i,\cdot}$, then we can take $\lambda = A\sigma_\epsilon\sqrt{\log p/n}$ and establish a slightly better rate of convergence.

### A.4. Proof of Theorem 1.

We write

$$V = \frac{Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j \cdot \sigma_e^2}{(Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j)^2} \quad \text{and} \quad \sqrt{V} = \frac{\sqrt{Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j \cdot \sigma_e^2}}{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j}.$$

Note that the following limiting result (50) shows that $Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j$ converges to a positive value in probability. From the equation (11), we have the following expression

$$\frac{1}{\sqrt{V}}(\widehat{\beta}_j - \beta_j) = \frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\epsilon}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} + B_\beta + B_b, \tag{45}$$

where $B_\beta$ and $B_b$ are the (scaled) bias terms defined as

$$B_\beta = \frac{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}(\widehat{\beta}_{-j}^{init} - \beta_{-j})}{\sqrt{Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j \cdot \sigma_e^2}} \quad \text{and} \quad B_b = \frac{Z_j^\intercal(\mathcal{P}^{(j)})^2 Xb}{\sqrt{Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j \cdot \sigma_e^2}}.$$

We decompose

$$\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\epsilon}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} = \frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}e}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} + \frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\Delta}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}$$

with $\Delta_i = \psi^\intercal H_{i,\cdot} - b^\intercal X_{i,\cdot}$ for $1 \leq i \leq n$. Since $e_i$ is Gaussian and independent of $X_{i,\cdot}$ and $Z_j$ is a function of $X$, we establish

$$\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}e}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} \mid X \sim N(0,1). \tag{46}$$

It follows from Lemma 2 that

$$\frac{1}{n}\mathbb{E}\|\Delta\|_2^2 = \mathbb{E}|\Delta_i|^2 = \phi^\intercal\left(\mathrm{I}_q - \Psi\Sigma_X^{-1}\Psi^\intercal\right)\phi \lesssim \frac{q\sqrt{\log p}}{1+\lambda_q^2(\Psi)}. \tag{47}$$

By Cauchy inequality, we have

$$\left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\Delta}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}\right| \leq \frac{1}{\sigma_e^2}\|\Delta\|_2.$$

Combined with (47), we establish that, with probability larger than $1-(\log p)^{-1/2}$,

$$\left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\Delta}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}\right| \lesssim \sqrt{\frac{nq\log p}{1+\lambda_q^2(\Psi)}}. \tag{48}$$

If $\lambda_q^2(\Psi) \gg \max\{1, qn\log p\}$, we combine (46) and (48) and establish

$$\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\epsilon}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} \xrightarrow{d} N(0,1). \tag{49}$$

We establish in the following lemma that $B_b$ and $B_\beta$ converges to 0 in probability under certain model conditions. The proof of this lemma is presented in Section C.7. The proof relies on our established intermediary results: Corollary 2, Lemmas 1 and 2, and Propositions 4 and 5.

LEMMA 3. *Suppose that the conditions of Theorem 1 hold. Then we have*

$$\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}{\mathrm{Tr}[(\mathcal{P}^{(j)})^2]\sigma_j^2} \xrightarrow{p} 1 \tag{50}$$

$$\frac{Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j}{\mathrm{Tr}[(\mathcal{P}^{(j)})^4]\sigma_j^2} \xrightarrow{p} 1 \tag{51}$$

$$B_\beta \xrightarrow{p} 0 \qquad B_b \xrightarrow{p} 0. \tag{52}$$

By the decomposition (45) together with (49) and (52), we establish the limiting distribution in (24). The asymptotic expression of the variance V in (25) follows from (50) and (51).

**A.5. Verification of Assumption** (A2)**.** In the following, we verify the condition (A2) for a general class of models, whose proof can be found in Section C.8.

LEMMA 4. *Suppose that $\{\Psi_{\cdot,l}\}_{1\leq l\leq p}$ are generated as i.i.d. $q$-dimensional sub-Gaussian random vectors with mean zero and covariance $\Sigma_\Psi \in \mathbb{R}^{q\times q}$. If $q \ll p$, $\lambda_{\max}(\Sigma_\Psi)/\lambda_{\min}(\Sigma_\Psi) \leq C$ and $\|\phi\|_\infty/\lambda_{\min}(\Sigma_\Psi) \leq C$ for some positive constant $C > 0$, then with probability larger than $1-(\log p)^{2c}$, we have*

$$\lambda_q(\Psi) \geq \lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}\sqrt{\lambda_{\min}(\Sigma_\Psi)} \tag{53}$$

$$\max\{\|\Psi(\Omega_E)_{\cdot,j}\|_2, \|\Psi_j\|_2, \|\Psi_{-j}(\Omega_E)_{-j,j}\|_2, \|\phi\|_2\} \lesssim \sqrt{\lambda_{\max}(\Sigma_\Psi)}\cdot\sqrt{q}(\log p)^c, \tag{54}$$

*where $c > 0$ is a positive constant.*

The conclusion of Lemma 4 can be generalized to hold if a fixed proportion of the $p$ columns of $\Psi$ are i.i.d. sub-Gaussian in $\mathbb{R}^q$. This generalized result is stated in the following lemma, whose proof is presented in Section C.9:

LEMMA 5. *Suppose that there exists a set $A \subseteq \{1,2,\ldots,p\}$ such that $\{\Psi_{\cdot,l}\}_{l\in A}$ are generated as i.i.d sub-Gaussian random vector with mean zero and covariance $\Sigma_{\Psi} \in \mathbb{R}^{q\times q}$ and $\{\Psi_{\cdot,l}\}_{l\in A^c}$ are generated as independent $q$-dimensional sub-Gaussian random vectors with sub-Gaussian norm $C_1$. If $\max\{C_1, \lambda_{\max}(\Sigma_{\Psi})\}/\lambda_{\min}(\Sigma_{\Psi}) \leq C$, $\|\psi\|_{\infty}/\lambda_{\min}(\Sigma_{\Psi}) \leq C$ and $\max\{C_1, \lambda_{\max}(\Sigma_{\Psi})\} \leq C$ for some positive constant $C > 0$ and $|A|$ satisfies*

$$|A| \gg q \quad \textit{and} \quad |A| \gg \max\left\{\sqrt{\frac{qp}{n}}(\log p)^{3/4}, \sqrt{qn\log p}, q^{3/2}(\log p)^{3/4}\right\}, \tag{55}$$

*then the assumption* (A2) *holds with probability larger than $1-(\log p)^{2c}$.*

## APPENDIX B: PROOF OF PROPOSITION 1

We express the hidden confounding model as

$$X_{n\times p} = D_{n\times p} + E_{n\times p} \quad \text{with} \quad D_{n\times p} = H_{n\times q}\Psi_{q\times p}. \tag{56}$$

For a given $q$, a natural way to estimate $\Psi$ and $H$ is to solve the optimization problem $\arg\min_{H\in\mathbb{R}^{n\times q},\Psi\in\mathbb{R}^{q\times p}} \|X - H\Psi\|_F^2$, where $\|\cdot\|_F$ denotes the matrix Frobenius norm. Since the solution of this optimization problem is not unique, we introduce an additional constraint $H^{\intercal}H/n = \mathrm{I}_q$ for the parameter identification. Then the minimizer is defined as

$$\begin{aligned}(\widetilde{H}, \widetilde{\Psi}) &= \underset{H\in\mathbb{R}^{n\times q},\Psi\in\mathbb{R}^{q\times p},H^{\intercal}H/n=\mathrm{I}_q}{\arg\min} \|X - H\Psi\|_F^2 \\ &= \underset{H\in\mathbb{R}^{n\times q},\Psi\in\mathbb{R}^{q\times p},H^{\intercal}H/n=\mathrm{I}_q}{\arg\min} -2\mathrm{Tr}(\Psi^{\intercal}H^{\intercal}X) + n\mathrm{Tr}(\Psi^{\intercal}\Psi).\end{aligned}$$

We compute the derivative of $-2\mathrm{Tr}(\Psi^{\intercal}H^{\intercal}X) + n\mathrm{Tr}(\Psi^{\intercal}\Psi)$ with respect to $\Psi$ and set it to be zero. Then we obtain the solution

$$\frac{1}{n}\widetilde{H}^{\intercal}X = \widetilde{\Psi} \quad \text{with} \quad \widetilde{H} = \underset{H\in\mathbb{R}^{n\times q},H^{\intercal}H/n=\mathrm{I}_q}{\arg\max} \mathrm{Tr}(H^{\intercal}XX^{\intercal}H). \tag{57}$$

That is, the columns of $\widetilde{H} \in \mathbb{R}^{n\times q}$ are $\sqrt{n}$ times the first $q$ eigenvectors, corresponding to the top $q$ eigenvalues of $XX^{\intercal} \in \mathbb{R}^{n\times n}$. Then the PCA adjusted covariates are defined as

$$\widetilde{X}^{\mathrm{PCA}} = X - \widetilde{D} \quad \text{with} \quad \widetilde{D} = \widetilde{H}\widetilde{\Psi}.$$

That is, we remove from $X$ the eigen-decomposition corresponding to the top $q$ eigenvalues, which is denoted as $\widetilde{D}$. Define $R = D - \widetilde{D}$. Then we have $\widetilde{X}^{\mathrm{PCA}} = R + E$ and

$$\frac{1}{n}(\widetilde{X}^{\mathrm{PCA}})^{\intercal}\widetilde{X}^{\mathrm{PCA}} - \Sigma_E = \left(\frac{1}{n}E^{\intercal}E - \Sigma_E\right) + \frac{1}{n}R^{\intercal}E + \frac{1}{n}E^{\intercal}R + \frac{1}{n}R^{\intercal}R.$$

We further have

$$\begin{aligned}&\min_{\|\omega_{\mathcal{T}^c}\|_1 \leq CM\cdot\|\omega_{\mathcal{T}}\|_1, \|\omega\|_2=1} \omega^{\intercal}\left(\frac{1}{n}(\widetilde{X}^{\mathrm{PCA}})^{\intercal}\widetilde{X}^{\mathrm{PCA}} - \Sigma_E\right)\omega \\ &\geq \min_{\|\omega_{\mathcal{T}^c}\|_1 \leq CM\cdot\|\omega_{\mathcal{T}}\|_1, \|\omega\|_2=1} \omega^{\intercal}\left(\frac{1}{n}E^{\intercal}E - \Sigma_E\right)\omega \qquad (58) \\ &\quad - \max_{\|\omega_{\mathcal{T}^c}\|_1 \leq CM\cdot\|\omega_{\mathcal{T}}\|_1, \|\omega\|_2=1} \omega^{\intercal}\frac{2}{n}R^{\intercal}E\omega - \max_{\|\omega_{\mathcal{T}^c}\|_1 \leq CM\cdot\|\omega_{\mathcal{T}}\|_1, \|\omega\|_2=1} \omega^{\intercal}\frac{1}{n}R^{\intercal}R\omega.\end{aligned}$$

In the following, we shall control the three terms on the right-hand-side of (58).

Note that Theorem 1.6 in [66] (with $k_0$ in this theorem taken as $CM$) implies that, if

$$n \gtrsim M^2\frac{k\log p}{n},$$

then with probability larger than $1-p^{-c}$ for some positive constant $c>0$,

$$\max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\left|\sqrt{\frac{\omega^\intercal\frac{1}{n}E^\intercal E\omega}{\omega^\intercal\Sigma_E\omega}}-1\right|\leq 0.1.$$

That is, there exists a positive constant $C'>0$ such that

$$\min_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\omega^\intercal\frac{1}{n}E^\intercal E\omega\geq 0.9\cdot\lambda_{\min}(\Sigma_E), \tag{59}$$

and

$$\max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\omega^\intercal\frac{1}{n}E^\intercal E\omega\leq 1.1\cdot\lambda_{\max}(\Sigma_E). \tag{60}$$

Now we turn to $\omega^\intercal\frac{1}{n}R^\intercal R\omega$. Fix $\mathcal{T}\subseteq[p]$ with $|\mathcal{T}|\leq k$. Then we have

$$\begin{aligned}\max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\omega^\intercal\frac{1}{n}R^\intercal R\omega &\leq \max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\max_{1\leq l\leq n}(\sum_{j=1}^p R_{l,j}\omega_j)^2\\ &\leq \max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}(\|R\|_\infty\|\omega\|_1)^2\\ &\leq \max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}(\|R\|_\infty(1+CM)\|\omega_{\mathcal{T}}\|_1)^2\\ &\leq \max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}(\|R\|_\infty(1+CM)\sqrt{k}\|\omega_{\mathcal{T}}\|_2)^2\\ &\lesssim M^2\cdot k\|R\|_\infty^2.\end{aligned} \tag{61}$$

By combining (60) and (61), we establish

$$\begin{aligned}&\max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\omega^\intercal\frac{1}{n}R^\intercal E\omega\\ &\leq \max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\sqrt{\frac{1}{n}\omega^\intercal R^\intercal R\omega}\cdot\sqrt{\frac{1}{n}\omega^\intercal E^\intercal E\omega}\\ &\leq\sqrt{\max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\frac{1}{n}\omega^\intercal R^\intercal R\omega}\cdot\sqrt{\max_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\frac{1}{n}\omega^\intercal E^\intercal E\omega}\\ &\lesssim\sqrt{M^2\cdot k\|R\|_\infty^2}.\end{aligned} \tag{62}$$

By the decomposition in (58) and the bounds in (59), (61) and (62), we establish

$$\begin{aligned}&\min_{\|\omega_{\mathcal{T}^c}\|_1\leq CM\cdot\|\omega_{\mathcal{T}}\|_1,\|\omega\|_2=1}\omega^\intercal\left(\frac{1}{n}(\widetilde{X}^{\rm PCA})^\intercal\widetilde{X}^{\rm PCA}-\Sigma_E\right)\omega\\ &\geq 0.9\cdot\lambda_{\min}(\Sigma_E)-C\sqrt{M^2\cdot k\|R\|_\infty^2}-CM^2\cdot k\|R\|_\infty^2,\end{aligned}$$

where $C$ is a positive constant independent of $n$ and $p$. If

$$n\gtrsim M^2\cdot\frac{k\log p}{n}\quad\text{and}\quad M\cdot\sqrt{k}\|R\|_\infty\to 0, \tag{63}$$

we establish that, for a sufficiently large $n$, there exists a small positive constant $0<c<0.9$ independent of $n$ and $p$ such that

$$\text{RE}\left(\frac{1}{n}(\widetilde{X}^{\rm PCA})^\intercal\widetilde{X}^{\rm PCA}\right)\geq c\lambda_{\min}(\Sigma_E).$$

By the Weyl's inequality for singular values,

$$|\lambda_l(X) - \lambda_l(D)| = |\lambda_l(D+E) - \lambda_l(D)| \leq \|E\|_2 \quad \text{for} \quad 1 \leq l \leq p.$$

We then apply Theorem 5.39 of [57] and establish that, with probability larger than $1 - p^{-c}$ for some positive constant $c > 0$,

$$|\lambda_l(X) - \lambda_l(D)| \leq \|E\|_2 \lesssim \sqrt{n} + \sqrt{p} \quad \text{for} \quad 1 \leq l \leq p.$$

Since $D$ is of rank $q$, then $\lambda_{q+1}(D) = 0$ and hence

$$|\lambda_{q+1}(X)| \lesssim \sqrt{n} + \sqrt{p} \quad \text{and} \quad \left|\lambda_{q+1}\left(\frac{1}{n}XX^{\intercal}\right)\right| \lesssim \max\left\{1, \frac{p}{n}\right\}. \tag{64}$$

Recall that $X = \sum_{j=1}^{m} \Lambda_{j,j} U_{\cdot,j} V_{\cdot,j}^{\intercal}$. For any $\omega \in \mathbb{R}^p$ and $\lfloor \rho m \rfloor \geq q+1$, we have

$$\begin{aligned}
\omega^{\intercal} X^{\intercal} \mathcal{Q}^2 X \omega &= \omega^{\intercal} \sum_{j=1}^{\lfloor \rho m \rfloor} \Lambda^2_{\lfloor \rho m \rfloor, \lfloor \rho m \rfloor} V_{\cdot,j} V_{\cdot,j}^{\intercal} \omega + \omega^{\intercal} \sum_{j=\lfloor \rho m \rfloor + 1}^{m} \Lambda^2_{j,j} V_{\cdot,j} V_{\cdot,j}^{\intercal} \omega \\
&\geq \omega^{\intercal} \sum_{j=q+1}^{\lfloor \rho m \rfloor} \Lambda^2_{\lfloor \rho m \rfloor, \lfloor \rho m \rfloor} V_{\cdot,j} V_{\cdot,j}^{\intercal} \omega + \omega^{\intercal} \sum_{j=\lfloor \rho m \rfloor + 1}^{m} \Lambda^2_{j,j} V_{\cdot,j} V_{\cdot,j}^{\intercal} \omega \\
&\geq \frac{\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n} XX^{\intercal})}{\lambda_{q+1}(\frac{1}{n} XX^{\intercal})} \left( \omega^{\intercal} \sum_{j=q+1}^{\lfloor \rho m \rfloor} \Lambda^2_{j,j} V_{\cdot,j} V_{\cdot,j}^{\intercal} \omega + \omega^{\intercal} \sum_{j=\lfloor \rho m \rfloor + 1}^{m} \Lambda^2_{j,j} V_{\cdot,j} V_{\cdot,j}^{\intercal} \omega \right).
\end{aligned}$$

If $\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n} XX^{\intercal}) \geq c \max\{1, p/n\}$, together with (64), we establish that, there exists some positive constant $c' > 0$ such that, with probability larger than $1 - p^{-c}$,

$$\frac{\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n} XX^{\intercal})}{\lambda_{q+1}(\frac{1}{n} XX^{\intercal})} \geq c'.$$

This leads to

$$\omega^{\intercal} X^{\intercal} \mathcal{Q}^2 X \omega \gtrsim \omega^{\intercal} (\widetilde{X}^{\text{PCA}})^{\intercal} \widetilde{X}^{\text{PCA}} \omega$$

for any $\omega \in \mathbb{R}^p$ and hence with probability larger than $1 - p^{-c}$,

$$\text{RE}\left(\frac{1}{n} X^{\intercal} \mathcal{Q}^2 X\right) \gtrsim \text{RE}\left(\frac{1}{n} (\widetilde{X}^{\text{PCA}})^{\intercal} \widetilde{X}^{\text{PCA}}\right).$$

To complete the proof, we shall apply the following lemma to verify the dimension condition (63). The proof of the following lemma is presented at Section B.2.

LEMMA 6. *Suppose that assumptions (A1) and (A3) hold, $H_{i,\cdot}$ is a sub-Gaussian random vector, $q + \log p \lesssim \sqrt{n}$, $k = \|\beta\|_0$ satisfies $kq^2 \log p \log n / n \to 0$. The loading matrix $\Psi \in \mathbb{R}^{q \times p}$ satisfies $\max_{1 \leq i \leq q, 1 \leq j \leq p} |\Psi_{i,j}| \lesssim \sqrt{\log(qp)}$, $\lambda_1(\Psi)/\lambda_q(\Psi) \leq C$ for some positive constant $C > 0$ and (23). Then with probability larger than $1 - p^{-c} - \exp(-cn)$ for some positive constant $c > 0$,*

$$\begin{aligned}
\|R\|_{\infty} \lesssim & \sqrt{\frac{q \log p}{n}} \sqrt{q \log(qn)} + \frac{q^{\frac{9}{2}} (\log N)^{\frac{7}{2}}}{\min\{n, p\}} \cdot \left(\frac{p}{\lambda_q^2(\Psi)}\right)^2 \sqrt{q \log(qp)} \\
& + \left(\sqrt{\frac{\log p}{n}} + \frac{q \log N}{\sqrt{p}}\right) \cdot \frac{p}{\lambda_q^2(\Psi)} \sqrt{q \log(qn)}.
\end{aligned} \tag{65}$$

Hence the dimension condition

$$\frac{M^2 \cdot kq^2 \log p \log n}{n} \to 0$$

together with (23) implies (63).

Furthermore, in Section B.1, we provide theoretical justification on the lower bound $\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n}XX^{\intercal})$.

### B.1. Lower bounds for $\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n}XX^{\intercal})$.

LEMMA 7. *Suppose that assumptions (A1) and (A3) hold and $H_{i,\cdot}$ is a sub-Gaussian random vector. With probability larger than $1 - p^{-c}$ for some positive constant $c > 0$, if either of the following two assumptions hold for $Z_{i,\cdot} = \Sigma_X^{-1/2} X_{i,\cdot}$:*

1. *$p/n \to c^* \in [0, \infty)$ and $\frac{1}{p}\left(Z_{i,\cdot}^{\intercal} A Z_{i,\cdot} - \mathrm{Tr}(A)\right) \xrightarrow{p} 0$ as $p \to \infty$ for all sequences of complex matrices $A \in \mathbb{R}^{p \times p}$ with uniformly bounded spectral norms $\|A\|_2$.*
2. *$p/n \to \infty$ and the entries of $Z_{i,\cdot}$ are independent.*

*then $\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n}XX^{\intercal}) \gtrsim \max\{1, p/n\}$ for $n$ sufficiently large.*

The condition $\frac{1}{p}\left(Z_{i,\cdot}^{\intercal} A Z_{i,\cdot} - \mathrm{Tr}(A)\right) \xrightarrow{p} 0$ is implied by the forth order moment condition: for $1 \le i \le n$,

$$E[Z_{i,j_1} Z_{i,j_2} Z_{i,j_3} Z_{i,j_4}] = 0 \text{ for all } j_1 \notin \{j_2, j_3, j_4\}. \tag{66}$$

The moment condition (66) is substantially weaker than assuming independent entries of $Z_{i,\cdot}$. Both conditions 1. and 2. are imposed only for technical reasons so that we can directly apply the lower bounds for the median (or smallest) singular values established in [63, 57, 51].

We now apply (66) to establish $\frac{1}{p}\left(Z_{i,\cdot}^{\intercal} A Z_{i,\cdot} - \mathrm{Tr}(A)\right) \xrightarrow{p} 0$. Note that

$$\begin{aligned}
\mathbb{E}\left|Z_{i,\cdot}^{\intercal} A Z_{i,\cdot} - \mathrm{Tr}(A)\right|^2 &= \mathbb{E}\left|Z_{i,\cdot}^{\intercal} A Z_{i,\cdot}\right|^2 - |\mathrm{Tr}(A)|^2 \\
&\le \sum_{1 \le j \ne l \le p} \mathbb{E} Z_{i,j}^2 Z_{i,l}^2 |A_{j,l}|^2 + \sum_{1 \le j \ne l \le p} \mathbb{E} Z_{i,j}^2 Z_{i,l}^2 |A_{j,l}||A_{l,j}| \\
&\lesssim \sum_{1 \le j \ne l \le p} |A_{j,l}|^2 \lesssim p
\end{aligned}$$

where the first equality uses that $\mathbb{E}[Z_{i,\cdot}^{T} A Z_{i.}] = \mathrm{Tr}(A)$, the first inequality follows from (66) and the last inequality follows from the bounded spectrum norm condition. Then we apply Markov's inequality to establish $\frac{1}{p}\left(Z_{i,\cdot}^{\intercal} A Z_{i,\cdot} - \mathrm{Tr}(A)\right) \xrightarrow{p} 0$ as $p \to \infty$.

We now present the proof of Lemma 7. With $Z = X\Sigma_X^{-\frac{1}{2}}$, we have

$$\lambda_{\min}(ZZ^{\intercal}) = \lambda_{\min}\left(X\Sigma_X^{-1}X^{\intercal}\right) \le \frac{1}{\lambda_{\min}(\Sigma_X)} \lambda_{\min}(XX^{\intercal}). \tag{67}$$

Note that $Z_{i,\cdot} = \Sigma_X^{-\frac{1}{2}} \Psi^{\intercal} H_{i\cdot} + \Sigma_X^{-\frac{1}{2}} E_{i\cdot}$. For any $v \in \mathbb{R}^p$ and $\|v\|_2 \le 1$, the random variable $v^{\intercal} Z_{i,\cdot}$ has sub-Gaussian norm upper bounded by $C\left(\|v^{\intercal}\Sigma_X^{-\frac{1}{2}}\Psi^{\intercal}\|_2 + \|\Sigma_X^{-\frac{1}{2}} v\|_2\right)$ for some positive constant $C > 0$. By (134), we show that $v^{\intercal} Z_{i,\cdot}$ has a bounded sub-Gaussian norm and hence $Z_{i,\cdot}$ is sub-Gaussian.

We now establish the lower bound for $\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n}XX^{\intercal})$ by considering two cases.
**Case 1:** $p/n \to c_* \in (0,\infty)$. For any set $B \subseteq \mathbb{R}$, define $\mu_p(B) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}(\lambda_j \in B)$ where $\{\lambda_j\}_{1\leq j\leq p}$ are eigenvalues of $\frac{1}{n}ZZ^{\intercal}$. Let $\mu_{c_*}$ denote the Marchenko Pastur law: for any set $B \subseteq \mathbb{R}$,

$$\mu_{c_*}(B) = \begin{cases} (1-1/c_*)\cdot \mathbf{1}(0\in B) + \int_a^b \frac{\sqrt{(b-t)(t-a)}}{2\pi c_* t}\cdot \mathbf{1}(t\in B)dt & \text{if} \quad c_* > 1 \\ \int_a^b \frac{\sqrt{(b-t)(t-a)}}{2\pi c_* t}\cdot \mathbf{1}(t\in B)dt & \text{if} \quad 0 < c_* \leq 1 \end{cases}$$

where $a = (1-\sqrt{c_*})^2$ and $b = (1+\sqrt{c_*})^2$.

In the following, we shall apply Theorem 1 of [63] and establish

$$\mu_p \xrightarrow{d} \mu_{c_*} \quad \text{almost surely.} \tag{68}$$

Note that Theorem 1 of [63] holds under the condition that

$$\frac{1}{p}\left(Z_{i,\cdot}^{\intercal} A Z_{i,\cdot} - \mathrm{Tr}(A)\right) \xrightarrow{p} 0$$

as $p \to \infty$ for any sequence of complex matrices $A \in \mathbb{R}^{p\times p}$ with uniformly bounded spectral norms $\|A\|_2$.

We now apply (68). When $c_* \neq 1$, (68) implies that

$$\liminf_{n\to\infty} \lambda_{\min}\left(\frac{1}{n}ZZ^{\intercal}\right) \geq (1-\sqrt{c_*})^2 \quad \text{almost surely.} \tag{69}$$

When $c_* = 1$, we need to calculate the median (or more general quantiles) of the distribution with the density function $\frac{\sqrt{(4-t)t}}{2\pi t}$. For $\rho = 1/2$, the median is within the range between $0.65$ and $0.66$, which, together with (68), lead to

$$\liminf_{n\to\infty} \lambda_{\lfloor m/2 \rfloor}\left(\frac{1}{n}ZZ^{\intercal}\right) \geq 0.65 \quad \text{almost surely.} \tag{70}$$

We combine (67), (69) and (70) and show that there exists some constant $c > 0$ such that

$$\liminf_{n\to\infty} \lambda_{\lfloor m/2 \rfloor}\left(\frac{1}{n}XX^{\intercal}\right) \geq c\lambda_{\min}(\Sigma_X) \quad \text{almost surely.} \tag{71}$$

**Case 2:** $p/n \geq C$ **for some positive constant** $C > 0$ **and the entries of** $Z_{i,\cdot}$ **are independent.** Theorem 5.39 of [57] implies that with probability larger than $1-p^{-c}$, $\lambda_m(Z) \geq \sqrt{p} - C\sqrt{n} - \sqrt{\log p}$, where $C$ is the constant defined in [57] and independent of $n$ and $p$. Combined with (67), we establish that, with probability larger than $1-p^{-c}$,

$$\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n}XX^{\intercal}) \geq \lambda_m(\frac{1}{n}XX^{\intercal}) \gtrsim \frac{p}{n}\lambda_{\min}(\Sigma_X). \tag{72}$$

**B.2. Proof of Lemma 6.** We prove the lemma through a finite-sample analysis of the factor model (56). The proof idea follows from that in [1] and [2], who establish the limiting distribution for any single entry of the matrix $R = \widetilde{D} - M$; see Theorem 3 in [1] for details. In our following proof, the main difference is to establish the rate of convergence of $\|R\|_\infty$ using finite-sample concentration bounds. We also relax the strong factor assumption $\lambda_q(\Psi) \asymp \sqrt{p}$ in [2] to the weaker condition (23).

Define $\widehat{\Lambda}^2 \in \mathbb{R}^{q\times q}$ to be the diagonal matrix consisting of the top $q$ eigenvalues of the matrix $\frac{1}{np}XX^{\intercal}$. Define

$$\mathcal{O} = (\Psi\Psi^{\intercal}/p)(H^{\intercal}\widetilde{H}/n)\widehat{\Lambda}^{-2} \in \mathbb{R}^{q\times q}. \tag{73}$$

Define $N=\max\{n,p\}$. Define the events

$$\mathcal{G}_1=\left\{\|\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}H_{i,\cdot}^{\intercal}-\mathrm{I}\|_2\lesssim\sqrt{\frac{q+\log p}{n}}\right\}$$

$$\mathcal{G}_2=\left\{\max_{1\le i\le n}\|H_{i,\cdot}\|_2\lesssim\sqrt{q\log(nq)}\right\}$$

$$\mathcal{G}_3=\left\{\max_{1\le t\le n}\|\Psi^{\intercal}H_{t,\cdot}/p\|_2\lesssim\frac{\sqrt{q}\sqrt{\log(pq)}\sqrt{\log(np)}}{\sqrt{p}}\right\}$$

$$\mathcal{G}_4=\left\{\max_{1\le t\le n}\max_{1\le j\le q}\frac{1}{\|\Psi_{j,\cdot}\|_2}\left|\Psi_{j,\cdot}^{\intercal}E_{t,\cdot}\right|\lesssim\sqrt{\log N}\right\}$$

$$\mathcal{G}_5=\left\{\max_{1\le i\le n}E_{i,\cdot}^{\intercal}E_{i,\cdot}/p\lesssim\log(np)\right\}$$

$$\mathcal{G}_6=\left\{\max_{1\le t\ne i\le n}\left|E_{i,\cdot}^{\intercal}E_{t,\cdot}/p\right|\lesssim\frac{\sqrt{\log p}\sqrt{\log(np)}}{\sqrt{p}}\right\}$$

$$\mathcal{G}_7=\left\{\frac{\|H^{\intercal}E\Psi^{\intercal}\|_2}{np}=\left\|\frac{1}{n}\sum_{i=1}^{n}H_{i\cdot}\frac{1}{p}E_{i\cdot}^{\intercal}\Psi^{\intercal}\right\|_2\lesssim\sqrt{\frac{q+\log p}{n}}\cdot\frac{\lambda_{\max}(\Psi)}{p}\right\}$$

$$\mathcal{G}_8=\left\{\|H\|_2\lesssim\sqrt{n},\|E\|_2\lesssim\sqrt{n}+\sqrt{p}\right\}$$

$$\mathcal{G}_9=\left\{\frac{\|H^{\intercal}E\|_2}{np}=\left\|\frac{1}{n}\sum_{i=1}^{n}H_{i\cdot}\frac{1}{p}E_{i\cdot}^{\intercal}\right\|_2\lesssim\frac{1}{p}+\frac{1}{\sqrt{np}}\right\}$$

$$\mathcal{G}_{10}=\left\{c\frac{\lambda_q(\Psi)}{\sqrt{p}}\le\lambda_{\min}(\widehat{\Lambda})\le\lambda_1(\widehat{\Lambda})\le C\frac{\lambda_1(\Psi)}{\sqrt{p}},\ \lambda_{\max}(\mathcal{O})\le C\right\}$$

$$\mathcal{G}_{11}=\left\{\max_{1\le j,l\le p}\left|\frac{1}{n}\sum_{t=1}^{n}E_{t,j}E_{t,l}-(\Sigma_E)_{j,l}\right|\lesssim\sqrt{\frac{\log p}{n}}\right\}$$

$$\mathcal{G}_{12}=\left\{\max_{1\le j\le q,1\le l\le p}\left|\frac{1}{n}\sum_{t=1}^{n}H_{t,j}E_{t,l}\right|\lesssim\sqrt{\frac{\log p}{n}}\right\}$$

where $C>0$ and $c>0$ are some positive constants. Define

$$\mathcal{G}=\cap_{j=1}^{12}\mathcal{G}_j.$$

On the event $\mathcal{G}_4$, we have

$$\max_{1\le t\le n}\|\frac{1}{p}\Psi E_{t,\cdot}\|_2\lesssim\sqrt{q}\cdot\max_{1\le j\le q}\frac{\|\Psi_{j,\cdot}\|_2\sqrt{\log N}}{p}\lesssim\frac{q\log N}{\sqrt{p}}.\tag{74}$$

The following lemma shows that the event $\mathcal{G}$ happens with a high probability, whose proof can be found in Section B.5.

LEMMA 8. *Suppose that the conditions of Lemma 6 hold, then we have*

$$\mathbb{P}(\mathcal{G})\ge 1-p^{-c}-\exp(-cn)\tag{75}$$

*for some positive constant* $c>0$.

The following lemma characterizes the accuracy of the loading estimation, which can be viewed as the finite sample version of Theorem 1 in [2] and Theorem 1 in [1]. The proof of the following lemma can be found in Section B.3.

LEMMA 9. *On the event $\mathcal{G}$,*

$$\max_{1\leq t\leq n}\|\widetilde{H}_{t,\cdot}-\mathcal{O}^{\intercal}H_{t,\cdot}\|_2\lesssim\frac{p}{\lambda_q^2(\Psi)}\left(\frac{q^2(\log N)^{3/2}}{\sqrt{p}}+\sqrt{\frac{q\log N}{n}}\right)\tag{76}$$

*with $N=\max\{n,p\}$. Furthermore, with probability larger than $1-n^{-c}-p^{-c}$ for some constant $c>0$,*

$$\left|\widetilde{H}_{t,\cdot}-\mathcal{O}^{\intercal}H_{t,\cdot}-\widehat{\Lambda}^{-2}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}\right|\lesssim\left(\frac{p}{\lambda_q^2(\Psi)}\right)^2\frac{q^{\frac{7}{2}}(\log N)^3}{\min\{n,p\}}\tag{77}$$

*and*

(78)

$$\left\|\widehat{\Lambda}^{-2}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}\right\|_2\leq\|\widehat{\Lambda}^{-2}\|_2\cdot\|\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}H_{i,\cdot}^{\intercal}\|_2\cdot\|\frac{1}{p}\Psi E_{t,\cdot}\|_2\lesssim\frac{p}{\lambda_q^2(\Psi)}\cdot\frac{q\log N}{\sqrt{p}}.$$

The following lemma characterizes the accuracy of the loading estimation, which can be viewed as the finite sample version of Theorem 2 in [1]. The proof of the following lemma can be found in Section B.4.

LEMMA 10. *On the event $\mathcal{G}$,*

(79)

$$\max_{1\leq l\leq p}\left\|\widetilde{\Psi}_{\cdot,l}-\mathcal{O}^{-1}\Psi_{\cdot,l}\right\|\lesssim\frac{q^{\frac{9}{2}}(\log N)^{\frac{7}{2}}}{\min\{n,p\}}\cdot\left(\frac{p}{\lambda_q^2(\Psi)}\right)^2+\left(\sqrt{\frac{\log p}{n}}+\frac{1}{\sqrt{p}}\right)\cdot\frac{p}{\lambda_q^2(\Psi)}+\sqrt{\frac{q\log p}{n}}.$$

For $1\leq t\leq n$ and $1\leq l\leq p$, we have the the following decomposition for $\widetilde{D}_{t,l}-M_{t,l}$

(80)

$$\begin{aligned}
&\widetilde{H}_{t,\cdot}^{\intercal}\widetilde{\Psi}_{\cdot,l}-H_{t,\cdot}^{\intercal}\Psi_{\cdot,l}\\
&=\widetilde{H}_{t,\cdot}^{\intercal}\widetilde{\Psi}_{\cdot,l}-(\mathcal{O}^{\intercal}H_{t,\cdot})^{\intercal}\mathcal{O}^{-1}\Psi_{\cdot,l}\\
&=(\widetilde{H}_{t,\cdot}-\mathcal{O}^{\intercal}H_{t,\cdot})^{\intercal}\widetilde{\Psi}_{\cdot,l}+(\mathcal{O}^{\intercal}H_{t,\cdot})^{\intercal}(\widetilde{\Psi}_{\cdot,l}-\mathcal{O}^{-1}\Psi_{\cdot,l})\\
&=(\widetilde{H}_{t,\cdot}-\mathcal{O}^{\intercal}H_{t,\cdot})^{\intercal}\mathcal{O}^{-1}\Psi_{\cdot,l}+(\mathcal{O}^{\intercal}H_{t,\cdot})^{\intercal}(\widetilde{\Psi}_{\cdot,l}-\mathcal{O}^{-1}\Psi_{\cdot,l})+(\widetilde{H}_{t,\cdot}-\mathcal{O}^{\intercal}H_{t,\cdot})^{\intercal}(\widetilde{\Psi}_{\cdot,l}-\mathcal{O}^{-1}\Psi_{\cdot,l}).
\end{aligned}$$

On the event $\mathcal{G}_2\cap\mathcal{G}_{10}$, we have

$$\|\mathcal{O}^{-1}\Psi_{\cdot,l}\|_2\lesssim\sqrt{q\log(qp)}\quad\text{and}\quad\|\mathcal{O}^{\intercal}H_{t,\cdot}\|_2\lesssim\sqrt{q\log(nq)}.$$

Note that

$$\|R\|_\infty=\max_{1\leq t\leq n,1\leq l\leq p}|\|\widetilde{H}_{t,\cdot}^{\intercal}\widetilde{\Psi}_{\cdot,l}-H_{t,\cdot}^{\intercal}\Psi_{\cdot,l}\|_2.$$

By applying Lemmas 9 and 10 to the decomposition (80), we establish that (65) holds on the event $\mathcal{G}$.

**B.3. Proof of Lemma 9.** Recall that $\widehat{\Lambda}^2 \in \mathbb{R}^{q\times q}$ denotes the diagonal matrix consisting of the top $q$ eigenvalues of the matrix $\frac{1}{np}XX^{\intercal}$. By the definition of $\widetilde{H}$ in (57), we have

$$\widetilde{H} = \frac{1}{np}XX^{\intercal}\widetilde{H}\widehat{\Lambda}^{-2}.$$

With the above expression, we establish the following decomposition of $\widetilde{H}_{t,\cdot} - \mathcal{O}^{\intercal}H_{t,\cdot} \in \mathbb{R}^q$ for $1 \le t \le n$,

$$\begin{aligned}
\widetilde{H}_{t,\cdot} - \mathcal{O}^{\intercal}H_{t,\cdot} &= \frac{1}{np}\widehat{\Lambda}^{-2}\widetilde{H}^{\intercal}XX_{t,\cdot} - \mathcal{O}^{\intercal}H_{t,\cdot} \\
&= \frac{1}{np}\widehat{\Lambda}^{-2}\widetilde{H}^{\intercal}(H\Psi + E)(\Psi^{\intercal}H_{t,\cdot} + E_{t,\cdot}) - \mathcal{O}^{\intercal}H_{t,\cdot} \\
&= \frac{1}{np}\widehat{\Lambda}^{-2}\widetilde{H}^{\intercal}H\Psi E_{t,\cdot} + \frac{1}{np}\widehat{\Lambda}^{-2}\widetilde{H}^{\intercal}E\Psi^{\intercal}H_{t,\cdot} + \frac{1}{np}\widehat{\Lambda}^{-2}\widetilde{H}^{\intercal}EE_{t,\cdot} \\
&= \widehat{\Lambda}^{-2}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot} + \frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot} + \frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\right).
\end{aligned}\tag{81}$$

Proof of (76). By (81), we have

$$\begin{aligned}
&\|\widetilde{H}_{t,\cdot} - \mathcal{O}^{\intercal}H_{t,\cdot}\|_2 \\
\le&\|\widehat{\Lambda}^{-2}\|_2\left(\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}\|_2 + \|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\|_2 + \|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\|_2\right).
\end{aligned}\tag{82}$$

We upper bound the three terms on the right hand side of (82) as

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}\|_2 &\le \frac{1}{n}\sum_{i=1}^{n}\|\widetilde{H}_{i,\cdot}\|_2|\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}| \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\widetilde{H}_{i,\cdot}\|_2^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}|^2};
\end{aligned}\tag{83}$$

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\|_2 &\le \frac{1}{n}\sum_{i=1}^{n}\|\widetilde{H}_{i,\cdot}\|_2|\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi H_{t,\cdot}| \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\widetilde{H}_{i,\cdot}\|_2^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi H_{t,\cdot}|^2};
\end{aligned}\tag{84}$$

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\|_2 &\le \frac{1}{n}\sum_{i=1}^{n}\|\widetilde{H}_{i,\cdot}\|_2|\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}| \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\widetilde{H}_{i,\cdot}\|_2^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}|^2}.
\end{aligned}\tag{85}$$

Note that

$$\max_{1\le t\le n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}|^2} \le \max_{1\le i\le n}\|H_{i,\cdot}\|_2\max_{1\le t\le n}\|\Psi E_{t,\cdot}/p\|_2,$$

$$\max_{1\leq t\leq n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi H_{t,\cdot}|^2}\leq \max_{1\leq t\leq n}\|H_{t,\cdot}\|_2\max_{1\leq i\leq n}\|\Psi E_{i,\cdot}/p\|_2.$$

Together with (74), we establish that, on the event $\mathcal{G}_2$,

$$\max\left\{\max_{1\leq t\leq n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}|^2},\max_{1\leq t\leq n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi H_{t,\cdot}|^2}\right\}\lesssim\frac{q^{\frac{3}{2}}(\log N)^{\frac{3}{2}}}{\sqrt{p}}.$$

Note that $\widetilde{H}^{\intercal}\widetilde{H}/n=\mathrm{I}$ implies $\frac{1}{n}\sum_{i=1}^{n}\|\widetilde{H}_{i,\cdot}\|_2^2=q$. Combined with (83) and (84), we establish that, on the event $\mathcal{G}$,

$$\max\left\{\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}\|_2\|\frac{1}{n},\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\|_2\right\}\lesssim\frac{q^2(\log N)^{\frac{3}{2}}}{\sqrt{p}}. \tag{86}$$

Note that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}|^2}=\sqrt{\frac{1}{n}\sum_{t\neq i}|\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}|^2+\frac{1}{n}|\frac{1}{p}E_{t,\cdot}^{\intercal}E_{t,\cdot}|^2}.$$

On the event $\mathcal{G}_5\cap\mathcal{G}_6$, we have

$$\max_{1\leq t\leq n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}|^2}\lesssim\sqrt{\frac{q\log p\log(np)}{p}+\frac{\log(np)}{n}}.$$

Combined with (85), we establish

$$\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\|_2\lesssim\frac{q\log N}{\sqrt{p}}+\sqrt{\frac{q\log N}{n}}. \tag{87}$$

Together with (86), (87) and the definition of $\mathcal{G}_{10}$, we apply the decomposition (82) and establish (76).

Proof of (77). We shall establish the bound by applying (81) and the bound (76). Note the following three decompositions

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}$$

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}$$

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}$$

By applying (81) and the above three decompositions, we establish

(88)

$$\begin{aligned}&\widehat{\Lambda}^2(\frac{1}{np}\widehat{\Lambda}^{-2}\widetilde{H}^{\intercal}XX_{t,\cdot}-\mathcal{O}^{\intercal}H_{t,\cdot}-\widehat{\Lambda}^{-2}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot})\\&=\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}+\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\\&+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}+\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\end{aligned}$$

Note that

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}\right| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}/p)^2}$$

and

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\right| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}/p)^2}.$$

On the event $\mathcal{G}$, we have (74) and then

$$\max_{1\leq t\leq n}\max_{1\leq i\leq n}\left|H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}/p\right| \leq \max_{1\leq i\leq n}\|H_{i,\cdot}\|_2\max_{1\leq t\leq n}\|\Psi E_{t,\cdot}/p\|_2 \lesssim \frac{(q\log N)^{3/2}}{\sqrt{p}}.$$

With the above three inequalities, we apply (76) and establish

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}\right| + \left|\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\right| \\ &\lesssim \frac{p}{\lambda_q^2(\Psi)}\frac{q^{\frac{7}{2}}(\log N)^3}{\sqrt{p}\sqrt{\min\{n,p\}}}\end{aligned} \tag{89}$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_{i,\cdot}^{\intercal}E_{t,\cdot}/p)^2}.$$

On the event $\mathcal{G}$, we apply (76) and (87) and establish

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot})\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\right| \leq \frac{p}{\lambda_q^2(\Psi)}\cdot\frac{q^3(\log N)^{\frac{5}{2}}}{\min\{n,p\}}. \tag{90}$$

We now turn to the upper bound for $\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}$ and first consider the setting $i\neq t$. Note that

$$\left\|\frac{1}{n}\sum_{i\neq t}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\right\|_2 \leq \sqrt{q}\max_{1\leq j\leq q}\left|\frac{1}{n}\sum_{i\neq t}H_{i,j}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\right|. \tag{91}$$

Conditioning on $H_{t,\cdot}$, the random variable $H_{i,j}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}$ is of zero mean and sub-exponential with sub-exponential norm upper bounded by $C\|\Psi^{\intercal}H_{t,\cdot}/p\|_2$. By Proposition 5.16 of [57], we establish

$$\mathbb{P}\left(\max_{1\leq j\leq q}\left|\frac{1}{n}\sum_{i\neq t}H_{i,j}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\right| \geq C\|\Psi^{\intercal}H_{t,\cdot}/p\|_2\sqrt{\frac{\log n}{n}} \mid H_{t,\cdot}\right) \leq n^{-c}.$$

Together with the definition of the event $\mathcal{G}_3$, we establish that, with probability larger than $(1-n^{-c})\cdot\mathbb{P}(\mathcal{G}_3)$,

$$\max_{1\leq j\leq q}\left|\frac{1}{n}\sum_{i\neq t}H_{i,j}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}\right| \lesssim \frac{\sqrt{q}(\log N)^{\frac{3}{2}}}{\sqrt{np}}. \tag{92}$$

On the event $\mathcal{G}_2 \cap \mathcal{G}_4$, we apply (74) and establish that for any $1 \leq t \leq n$,

$$\left\| \frac{1}{n} H_{t,\cdot} \frac{1}{p} E_{t,\cdot}^{\intercal} \Psi^{\intercal} H_{t,\cdot} \right\|_2 \leq \frac{1}{n} \|H_{t,\cdot}\|_2^2 \|\frac{1}{p} E_{t,\cdot}^{\intercal} \Psi^{\intercal}\|_2 \lesssim \frac{(q \log N)^2}{n\sqrt{p}}.$$

Together with (92), we establish

$$\left\| \frac{1}{n} \sum_{i=1}^{n} H_{i,\cdot} \frac{1}{p} E_{i,\cdot}^{\intercal} \Psi^{\intercal} H_{t,\cdot} \right\|_2 \lesssim \frac{q (\log N)^{\frac{3}{2}}}{\sqrt{np}} + \frac{(q \log N)^2}{n\sqrt{p}}. \tag{93}$$

We now consider the upper bound for $\mathcal{O}^{\intercal} \frac{1}{n} \sum_{i=1}^{n} H_{i,\cdot} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot}$ and consider the setting $i \neq t$. Note that

$$\left\| \frac{1}{n} \sum_{i \neq t} H_{i,\cdot} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot} \right\|_2 \leq \sqrt{q} \max_{1 \leq j \leq q} \left\| \frac{1}{n} \sum_{i \neq t} H_{i,j} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot} \right\|_2 \tag{94}$$

Conditioning on $E_{t,\cdot}$, the random variable $H_{i,j} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot}$ is of zero mean and sub-exponential with sub-exponential norm upper bounded by $C \|E_{t,\cdot}/p\|_2$. By Proposition 5.16 of [57], we establish

$$\mathbb{P}\left( \max_{1 \leq j \leq q} \left| \frac{1}{n} \sum_{i \neq t} H_{i,j} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot} \right| \geq C \|E_{t,\cdot}/p\|_2 \sqrt{\frac{\log n}{n}} \mid E_{t,\cdot} \right) \leq n^{-c}. \tag{95}$$

Together with the definition of $\mathcal{G}_5$, we show that, with probability larger than $(1 - n^{-c}) \cdot \mathbb{P}(\mathcal{G}_5)$,

$$\max_{1 \leq j \leq q} \left| \frac{1}{n} \sum_{i \neq t} H_{i,j} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot} \right| \lesssim \frac{\log N}{\sqrt{np}} \quad \text{and} \quad \left\| \frac{1}{n} \sum_{i \neq t} H_{i,\cdot} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot} \right\|_2 \lesssim \frac{\sqrt{q} \log N}{\sqrt{np}}. \tag{96}$$

On the event $\mathcal{G}_2 \cap \mathcal{G}_5$, we have

$$\left\| \frac{1}{n} H_{t,\cdot} \frac{1}{p} E_{t,\cdot}^{\intercal} E_{t,\cdot} \right\|_2 \leq \frac{1}{n} \|H_{t,\cdot}\|_2 \frac{1}{p} E_{t,\cdot}^{\intercal} E_{t,\cdot} \lesssim \frac{\sqrt{q} (\log N)^{\frac{3}{2}}}{n}.$$

Together with (96), we establish

$$\left\| \frac{1}{n} \sum_{i=1}^{n} H_{i,\cdot} \frac{1}{p} E_{i,\cdot}^{\intercal} E_{t,\cdot} \right\|_2 \lesssim \frac{\sqrt{q} \log N}{\sqrt{np}} + \frac{\sqrt{q} (\log N)^{\frac{3}{2}}}{n} \tag{97}$$

On the event $\mathcal{G}$, we apply the decomposition (88) with the error bounds (89),(90),(93), (97) and then establish (77). The upper bound in (78) follows from the definition of $\mathcal{G}_2$ and (74).

**B.4. Proof of Lemma 10.** By the definition of $\widetilde{\Psi}$ in (57), we now control the estimation error of $\widetilde{\Psi}_{\cdot,l} = \frac{1}{n} \widetilde{H}^{\intercal} X_{\cdot,l} \in \mathbb{R}^q$. We start with the following decomposition,

$$\begin{aligned} \widetilde{\Psi}_{\cdot,l} - \mathcal{O}^{-1} \Psi_{\cdot,l} &= \frac{1}{n} \widetilde{H}^{\intercal} (H \Psi_{\cdot,l} + E_{\cdot,l}) - \mathcal{O}^{-1} \Psi_{\cdot,l} \\ &= \frac{1}{n} \widetilde{H}^{\intercal} (H \Psi_{\cdot,l} + E_{\cdot,l}) - \mathcal{O}^{-1} \Psi_{\cdot,l} \\ &= \frac{1}{n} \widetilde{H}^{\intercal} \left( (H\mathcal{O} - \widetilde{H} + \widetilde{H}) \mathcal{O}^{-1} \Psi_{\cdot,l} + E_{\cdot,l} \right) - \mathcal{O}^{-1} \Psi_{\cdot,l} \\ &= \frac{1}{n} \widetilde{H}^{\intercal} (H - \widetilde{H} \mathcal{O}^{-1}) \Psi_{\cdot,l} + \frac{1}{n} \mathcal{O}^{\intercal} H^{\intercal} E_{\cdot,l} + \frac{1}{n} (\widetilde{H} - H\mathcal{O})^{\intercal} E_{\cdot,l}. \end{aligned} \tag{98}$$

To establish (79), we control all three terms on the right-hand-side of (98).

Control of $\frac{1}{n}\widetilde{H}^{\intercal}(H-\widetilde{H}\mathcal{O}^{-1})\Psi_{\cdot,l}$. Note that

$$
\begin{aligned}
&\frac{1}{n}\widetilde{H}^{\intercal}(H-\widetilde{H}\mathcal{O}^{-1})\Psi_{\cdot,l}\\
&=\frac{1}{n}\widetilde{H}^{\intercal}(H\mathcal{O}-\widetilde{H})\mathcal{O}^{-1}\Psi_{\cdot,l}\\
&=\frac{1}{n}\mathcal{O}^{\intercal}H^{\intercal}(H\mathcal{O}-\widetilde{H})\mathcal{O}^{-1}\Psi_{\cdot,l}+\frac{1}{n}(\widetilde{H}-H\mathcal{O})^{\intercal}(H\mathcal{O}-\widetilde{H})\mathcal{O}^{-1}\Psi_{\cdot,l}.
\end{aligned}
\tag{99}
$$

On the event $\mathcal{G}$, it follows from (76) that

$$
\begin{aligned}
\|\frac{1}{n}(\widetilde{H}-H\mathcal{O})^{\intercal}(H\mathcal{O}-\widetilde{H})\mathcal{O}^{-1}\Psi_{\cdot,l}\|_2 &\leq \frac{1}{n}\|\widetilde{H}-H\mathcal{O}\|_2^2\|\mathcal{O}^{-1}\Psi_{\cdot,l}\|_2\\
&\lesssim \frac{q^{\frac{9}{2}}(\log N)^{\frac{7}{2}}}{\min\{n,p\}}\cdot\left(\frac{p}{\lambda_q^2(\Psi)}\right)^2.
\end{aligned}
\tag{100}
$$

Since

$$
\frac{1}{n}\mathcal{O}^{\intercal}H^{\intercal}(H\mathcal{O}-\widetilde{H})\mathcal{O}^{-1}\Psi_{\cdot,l}=\frac{1}{n}\mathcal{O}^{\intercal}\sum_{t=1}^{n}H_{t,\cdot}(\mathcal{O}^{\intercal}H_{t,\cdot}-\widetilde{H}_{t,\cdot})^{\intercal}\mathcal{O}^{-1}\Psi_{\cdot,l},
$$

then on the event $\mathcal{G}$, we have

$$
\|\frac{1}{n}\mathcal{O}^{\intercal}H^{\intercal}(H\mathcal{O}-\widetilde{H})\mathcal{O}^{-1}\Psi_{\cdot,l}\|_2\lesssim\|\frac{1}{n}\sum_{t=1}^{n}H_{t,\cdot}(\mathcal{O}^{\intercal}H_{t,\cdot}-\widetilde{H}_{t,\cdot})^{\intercal}\|_2\sqrt{q\log N}.
\tag{101}
$$

In the following, we shall control

$$
\|\frac{1}{n}\sum_{t=1}^{n}H_{t,\cdot}(\mathcal{O}^{\intercal}H_{t,\cdot}-\widetilde{H}_{t,\cdot})^{\intercal}\|_2=\|\frac{1}{n}\sum_{t=1}^{n}(\mathcal{O}^{\intercal}H_{t,\cdot}-\widetilde{H}_{t,\cdot})H_{t,\cdot}^{\intercal}\|_2.
$$

It follows from (81) that $\widehat{\Lambda}^2\frac{1}{n}\sum_{t=1}^{n}(\mathcal{O}^{\intercal}H_{t,\cdot}-\widetilde{H}_{t,\cdot})H_{t,\cdot}^{\intercal}$ can be decomposed as

$$
\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}+\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}+\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\right)H_{t,\cdot}^{\intercal}\\
&=\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}H_{i,\cdot}^{\intercal}\right)\left(\frac{1}{np}\sum_{t=1}^{n}\Psi E_{t,\cdot}H_{t,\cdot}^{\intercal}\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}\right)\left(\frac{1}{n}\sum_{t=1}^{n}H_{t,\cdot}H_{t,\cdot}^{\intercal}\right)\\
&+\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\right)\left(\frac{1}{n}\sum_{t=1}^{n}E_{t,\cdot}H_{t,\cdot}^{\intercal}\right).
\end{aligned}
\tag{102}
$$

Note that

$$
\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}H_{i,\cdot}^{\intercal}=\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot}\right)H_{i,\cdot}^{\intercal}+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}H_{i,\cdot}^{\intercal}
$$

$$
\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}=\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot}\right)\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}
$$

$$
\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}=\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot}\right)\frac{1}{p}E_{i,\cdot}^{\intercal}+\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}
$$

On the event $\mathcal{G}$, we apply (76) and (74) and establish

$$\|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot}\right)H_{i,\cdot}^{\intercal}\|_2\lesssim\frac{q^{5/2}(\log N)^2}{\sqrt{\min\{n,p\}}}\cdot\frac{p}{\lambda_q^2(\Psi)},$$

$$\|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot}\right)\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}\|_2\lesssim\frac{q^{3}(\log N)^{\frac{5}{2}}}{\sqrt{\min\{n,p\}}\sqrt{p}}\cdot\frac{p}{\lambda_q^2(\Psi)},$$

$$\|\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{H}_{i,\cdot}-\mathcal{O}^{\intercal}H_{i,\cdot}\right)\frac{1}{p}E_{i,\cdot}^{\intercal}\|_2\lesssim\frac{q^{2}(\log N)^{\frac{5}{2}}}{\sqrt{\min\{n,p\}}\sqrt{p}}\cdot\frac{p}{\lambda_q^2(\Psi)}.$$

On the event $\mathcal{G}$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}H_{i,\cdot}^{\intercal}\right\|_2\lesssim 1+\frac{q^{5/2}(\log N)^2}{\sqrt{\min\{n,p\}}}\cdot\frac{p}{\lambda_q^2(\Psi)}. \tag{103}$$

$$\max\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}\right\|_2,\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\right\|_2\right\}\lesssim\frac{q^{3}(\log N)^{\frac{5}{2}}}{\sqrt{\min\{n,p\}}\sqrt{p}}\cdot\frac{p}{\lambda_q^2(\Psi)}. \tag{104}$$

Then on the event $\mathcal{G}$, we have established that

$$\|\widehat{\Lambda}^2\frac{1}{n}\sum_{t=1}^{n}(\mathcal{O}^{\intercal}H_{t,\cdot}-\widetilde{H}_{t,\cdot})H_{t,\cdot}^{\intercal}\|_2\lesssim\frac{q^{3}(\log N)^{\frac{5}{2}}}{\sqrt{\min\{n,p\}}\sqrt{p}}\cdot\frac{p}{\lambda_q^2(\Psi)}\cdot(1+\sqrt{p/n}).$$

Together with (100) and (101), we have

$$\left\|\frac{1}{n}\widetilde{H}^{\intercal}(H-\widetilde{H}\mathcal{O}^{-1})\Psi_{\cdot,l}\right\|_2\lesssim\frac{q^{\frac{9}{2}}(\log N)^{\frac{7}{2}}}{\min\{n,p\}}\cdot\left(\frac{p}{\lambda_q^2(\Psi)}\right)^2. \tag{105}$$

Control of $\frac{1}{n}\mathcal{O}^{\intercal}H^{\intercal}E_{\cdot,l}$ Note that

$$\frac{1}{n}\mathcal{O}^{\intercal}H^{\intercal}E_{\cdot,l}=\mathcal{O}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i,\cdot}E_{i,l}.$$

On the event $\mathcal{G}_{10}\cap\mathcal{G}_{12}$, we have

$$\left\|\frac{1}{n}\mathcal{O}^{\intercal}H^{\intercal}E_{\cdot,l}\right\|_2\lesssim\sqrt{\frac{q\log p}{n}}. \tag{106}$$

Control of $\frac{1}{n}(\widetilde{H}-H\mathcal{O})^{\intercal}E_{\cdot,l}$. It follows from (81) that the term $\widehat{\Lambda}^2\frac{1}{n}(\widetilde{H}-H\mathcal{O})^{\intercal}E_{\cdot,l}$ can be decomposed as

(107)
$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}H_{i,\cdot}^{\intercal}\Psi E_{t,\cdot}+\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}H_{t,\cdot}+\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}E_{t,\cdot}\right)E_{t,l}\\
&=\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}H_{i,\cdot}^{\intercal}\right)\frac{1}{p}\Psi\left(\frac{1}{n}\sum_{t=1}^{n}E_{t,\cdot}E_{t,l}-(\Sigma_E)_{\cdot,l}\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}H_{i,\cdot}^{\intercal}\right)\frac{1}{p}\Psi(\Sigma_E)_{\cdot,l}\\
&+\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}\right)\left(\frac{1}{n}\sum_{t=1}^{n}H_{t,\cdot}E_{t,l}\right)\\
&+\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\right)\left(\frac{1}{n}\sum_{t=1}^{n}E_{t,\cdot}E_{t,l}-(\Sigma_E)_{\cdot,l}\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\right)(\Sigma_E)_{\cdot,l}.
\end{aligned}$$

On the event $\mathcal{G}_{11} \cap \mathcal{G}_{12}$, together with the fact that $\lambda_{\max}(\Sigma_E) \leq C$ for some positive constant $C > 0$, we apply (103) and (104) and establish

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}H_{i,\cdot}^{\intercal}\right)\frac{1}{p}\Psi\left(\frac{1}{n}\sum_{t=1}^{n}E_{t,\cdot}E_{t,l}-(\Sigma_E)_{\cdot,l}\right)\right|+\left|\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}H_{i,\cdot}^{\intercal}\right)\frac{1}{p}\Psi(\Sigma_E)_{\cdot,l}\right|$$
$$\lesssim \frac{\lambda_1(\Psi)}{\sqrt{p}}\left(\sqrt{\frac{\log p}{n}}+\frac{1}{\sqrt{p}}\right)\left(1+\frac{q^{5/2}(\log N)^2}{\sqrt{\min\{n,p\}}}\cdot\frac{p}{\lambda_q^2(\Psi)}\right),$$

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\Psi^{\intercal}\right)\left(\frac{1}{n}\sum_{t=1}^{n}H_{t,\cdot}E_{t,l}\right)\right|\lesssim\frac{q^{\frac{7}{2}}(\log N)^{\frac{5}{2}}}{\sqrt{\min\{n,p\}}\sqrt{p}}\sqrt{\frac{\log p}{n}}\cdot\frac{p}{\lambda_q^2(\Psi)}.$$

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\right)\left(\frac{1}{n}\sum_{t=1}^{n}E_{t,\cdot}E_{t,l}-(\Sigma_E)_{\cdot,l}\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{H}_{i,\cdot}\frac{1}{p}E_{i,\cdot}^{\intercal}\right)(\Sigma_E)_{\cdot,l}\right|$$
$$\lesssim\frac{q^3(\log N)^{\frac{5}{2}}}{\sqrt{\min\{n,p\}}\sqrt{p}}\left(\sqrt{\frac{p\log p}{n}}+1\right)\cdot\frac{p}{\lambda_q^2(\Psi)}.$$

By the above bounds, we apply the decomposition (107) and establish

(108)
$$\left|\widehat{\Lambda}^2\frac{1}{n}(\widetilde{H}-H\mathcal{O})^{\intercal}E_{\cdot,l}\right|\lesssim\sqrt{\frac{\log p}{n}}+\frac{1}{\sqrt{p}}+\frac{q^3(\log N)^{\frac{5}{2}}}{\sqrt{\min\{n,p\}}\sqrt{p}}\left(\sqrt{\frac{p\log p}{n}}+1\right)\cdot\frac{p}{\lambda_q^2(\Psi)}.$$

A combination of (105), (108) and (106) leads to (79).

**B.5. Proof of Lemma 8.** Control of $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3$. By the equation (5.23) of [57], with probability larger than $1 - p^{-c}$, the event $\mathcal{G}_1$ holds. Note that

$$\max_{1\leq i\leq n}\|H_{i,\cdot}\|_2\leq\sqrt{q}\max_{1\leq i\leq n,1\leq j\leq q}|H_{i,j}|\quad\text{and}\quad\max_{1\leq t\leq n}\|\Psi^{\intercal}H_{t,\cdot}/p\|_2\leq\frac{1}{\sqrt{p}}\max_{1\leq t\leq n,1\leq j\leq p}\left|\Psi_j^{\intercal}H_{t,\cdot}\right|$$

Since $\{H_{i,\cdot}\}_{1\leq i\leq n}$ are i.i.d. sub-Gaussian vectors, with probability larger than $1-(p)^{-c}$,

$$\max_{1\leq i\leq n,1\leq j\leq q}|H_{i,j}|\lesssim\sqrt{\log(nq)}$$

and

$$\max_{1\leq t\leq n,1\leq j\leq p}\left|\Psi_j^{\intercal}H_{t,\cdot}\right|\lesssim\|\Psi_j\|_2\cdot\sqrt{\log(np)}\lesssim\sqrt{q}\sqrt{\log(pq)\log(np)}$$

Hence, we establish

$$\mathbb{P}\left(\mathcal{G}_1\cap\mathcal{G}_2\cap\mathcal{G}_3\right)\geq 1-p^{-c}.$$

Control of $\mathcal{G}_4 \cap \mathcal{G}_5 \cap \mathcal{G}_6$. For any $1 \leq j \leq q$, $\frac{1}{\|\Psi_{j,\cdot}\|_2}\Psi_{j,\cdot}^{\intercal}E_{t,\cdot}$ is sub-Gaussian random variable and this leads to $\mathbb{P}(\mathcal{G}_4) \geq 1-p^{-c}$. We also have $\mathbb{P}\left(\max_{t,j}|E_{t,j}|\lesssim\sqrt{\log(np)}\right)\geq 1-(np)^{-c}$, which leads to $\mathbb{P}(\mathcal{G}_5)\geq 1-p^{-c}$.

We fix $1 \leq t \leq n$ and consider $i \neq t$. Conditioning on $E_{t,\cdot}$, the random variable $E_{i,\cdot}^{\intercal}E_{t,\cdot}/p$ is a zero-mean sub-Gaussian random variable with sub-Gaussian norm $\|E_{t,\cdot}\|_2/p$. On the event $\mathcal{G}_5$, we establish

(109)
$$\mathbb{P}\left(\max_{i\neq t}|E_{i,\cdot}^{\intercal}E_{t,\cdot}/p|\lesssim\sqrt{\log p}\|E_{t,\cdot}\|_2/p\mid E_{t,\cdot}\right)\geq 1-p^{-c}.$$

Note that

(110)
$$
\begin{aligned}
&\mathbb{P}\left(\max_{i\neq t}|E_{i,\cdot}^{\intercal}E_{t,\cdot}/p|\lesssim\sqrt{\log p}\sqrt{p\log(np)}/p\right)\\
&\geq\mathbb{P}\left(\max_{i\neq t}|E_{i,\cdot}^{\intercal}E_{t,\cdot}/p|\lesssim\sqrt{\log p}\|E_{t,\cdot}\|_2/p,\ \|E_{t,\cdot}\|_2\lesssim\sqrt{p\log(np)}\right)\\
&\geq\int\mathbb{P}\left(\max_{i\neq t}|E_{i,\cdot}^{\intercal}E_{t,\cdot}/p|\lesssim\sqrt{\log p}\|E_{t,\cdot}\|_2/p\mid E_{t,\cdot}\right)\mathbf{1}(\|E_{t,\cdot}\|_2\lesssim\sqrt{p\log(np)})\mu(E_{t,\cdot})
\end{aligned}
$$

where $\mu(E_{t,\cdot})$ denotes the measure of $E_{t,\cdot}$. Combined with (109), we establish that, for a given $1\leq t\leq n$,

$$\mathbb{P}\left(\max_{i\neq t}|E_{i,\cdot}^{\intercal}E_{t,\cdot}/p|\lesssim\sqrt{\log p}\sqrt{p\log(np)}/p\right)\geq(1-p^{-c})\cdot\mathbb{P}(\mathcal{G}_5)\geq 1-p^{-c},$$

where $c>1$ is some positive constant. By applying another union bound, we establish $\mathbb{P}(\mathcal{G}_6)\geq 1-p^{-(c-1)}$. Hence, we establish

$$\mathbb{P}\left(\mathcal{G}_4\cap\mathcal{G}_5\cap\mathcal{G}_6\right)\geq 1-p^{-c}.$$

Control of $\mathcal{G}_7$. For any vector $u\in\mathbb{R}^q$ and $v\in\mathbb{R}^q$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}H_{i\cdot}\frac{1}{p}E_{i\cdot}^{\intercal}\Psi^{\intercal}\right\|_2=\sup_{u,v\in\mathbb{R}^q,\|u\|_2=1,\|v\|_2=1}u^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i\cdot}\frac{1}{p}E_{i\cdot}^{\intercal}\Psi^{\intercal}v$$

Since $H_{i\cdot}$ and $E_{i,\cdot}$ are sub-Gaussian random vectors, the random variable $u^{\intercal}H_{i\cdot}\frac{1}{p}E_{i\cdot}^{\intercal}\Psi^{\intercal}v$ is zero-mean with sub-exponential norm upper bounded by $C\frac{\|\Psi^{\intercal}v\|_2}{p}\lesssim\frac{\lambda_1(\Psi)}{p}$. We apply Corollary 5.17 of [57] and establish that, for $t\leq\sqrt{n}$,

$$\mathbb{P}\left(\left|u^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i\cdot}\frac{1}{p}E_{i\cdot}^{\intercal}\Psi^{\intercal}v\right|\gtrsim\frac{t}{\sqrt{n}}\cdot\frac{\lambda_1(\Psi)}{p}\right)\leq\exp(-ct^2).$$

We shall use $\mathcal{N}_q$ to denote the $\epsilon$-net of the unit ball in $\mathbb{R}^q$; see the definition of $\epsilon$-net in Definition 5.1 in [57]. Taking the union bound over all vectors $u,v\in\mathcal{N}_q$, we have

$$\mathbb{P}\left(\max_{u,v\in\mathcal{N}_q}\left|u^{\intercal}\frac{1}{n}\sum_{i=1}^{n}H_{i\cdot}\frac{1}{p}E_{i\cdot}^{\intercal}\Psi^{\intercal}v\right|\gtrsim\frac{t}{\sqrt{n}}\cdot\frac{\lambda_1(\Psi)}{p}\right)\leq|\mathcal{N}_q|^2\exp(-ct^2). \tag{111}$$

where $c>0$ is some positive constant. We choose $t^2=C\log(|\mathcal{N}_q|^2\cdot p)\leq\sqrt{n}$ for some positive constant $C>0$ such that $|\mathcal{N}_q|^2\exp(-ct^2)\leq p^{-c'}$ for some positive constant $c'>0$. By Lemmas 5.2 and 5.3 of [57], we take $|\mathcal{N}_q|^2=C^{2q}$ and apply (111) to establish that $\mathbb{P}(\mathcal{G}_7)\geq 1-p^{-c}$.

Control of $\mathcal{G}_8\cap\mathcal{G}_9$. By Theorem 5.39 of [57], we establish that $\mathbb{P}(\mathcal{G}_8)\geq 1-\exp(-c\min\{n,p\})$. Since $\|H^{\intercal}E\|_2\leq\|H\|_2\|E\|_2$, on the event $\mathcal{G}_8$, the event $\mathcal{G}_9$ holds. That is, we establish that $\mathbb{P}(\mathcal{G}_8\cap\mathcal{G}_9)\geq 1-\exp(-c\min\{n,p\})$.

Control of $\mathcal{G}_{10}$. We start with the decomposition

$$\frac{1}{np}XX^{\intercal}-\frac{1}{np}H\Psi\Psi^{\intercal}H^{\intercal}=\frac{1}{np}H\Psi E^{\intercal}+\frac{1}{np}E\Psi^{\intercal}H^{\intercal}+\frac{1}{np}EE^{\intercal}$$

On the event $\mathcal{G}_8$, we have

$$\|\frac{1}{np}E^{\intercal}E\|_2\leq\frac{1}{np}\|E\|_2^2\lesssim\frac{1}{n}+\frac{1}{p}$$

$$\left\|\frac{1}{np}H\Psi E^{\intercal}\right\|_2 \leq \frac{1}{np}\|H\|_2\|\Psi\|_2\|E\|_2 \lesssim \frac{1}{\sqrt{n}}+\frac{1}{\sqrt{p}}.$$

Then we have

$$\left\|\frac{1}{np}XX^{\intercal}-\frac{1}{np}H\Psi\Psi^{\intercal}H^{\intercal}\right\|_2 \lesssim \frac{1}{\sqrt{n}}+\frac{1}{\sqrt{p}}. \tag{112}$$

Note that the top $q$ eigenvalues of $\frac{1}{np}H\Psi\Psi^{\intercal}H^{\intercal}$ are the same as the top $q$ eigenvalues of $\frac{1}{np}\Psi^{\intercal}H^{\intercal}H\Psi$. We have

$$\frac{1}{np}\Psi^{\intercal}H^{\intercal}H\Psi = \frac{1}{p}\Psi^{\intercal}(H^{\intercal}H/n-\mathrm{I})\Psi+\frac{1}{p}\Psi^{\intercal}\Psi \tag{113}$$

On the event $\mathcal{G}_1$, we have

$$\|\frac{1}{p}\Psi^{\intercal}(H^{\intercal}H/n-\mathrm{I})\Psi\|_2 \lesssim \sqrt{\frac{q+\log p}{n}}\cdot\frac{\lambda_1^2(\Psi)}{p}.$$

Note that the top $q$ eigenvalues of $\frac{1}{p}\Psi^{\intercal}\Psi$ are the same as the top $q$ eigenvalues of $\frac{1}{p}\Psi\Psi^{\intercal}$. Hence, we have

$$\max_{1\leq i\leq q}\left|\lambda_i\left(\frac{1}{np}\Psi^{\intercal}H^{\intercal}H\Psi\right)-\lambda_i\left(\frac{1}{p}\Psi\Psi^{\intercal}\right)\right| \lesssim \sqrt{\frac{q+\log p}{n}}\cdot\frac{\lambda_1^2(\Psi)}{p}. \tag{114}$$

A combination of (112) and (114) leads to

$$\max_{1\leq i\leq q}\left|\lambda_i\left(\frac{1}{np}XX^{\intercal}\right)-\lambda_i\left(\frac{1}{p}\Psi\Psi^{\intercal}\right)\right| \lesssim \sqrt{\frac{q+\log p}{n}}\cdot\frac{\lambda_1^2(\Psi)}{p}+\frac{1}{\sqrt{p}}$$

By (23), there exists positive constants $C\geq c>0$ such that

$$c\frac{\lambda_q^2(\Psi)}{p}\leq \lambda_{\min}(\widehat{\Lambda}^2)\leq\lambda_{\max}(\widehat{\Lambda}^2)\leq C\frac{\lambda_1^2(\Psi)}{p}.$$

By the definition of $\mathcal{O}$ in (73), we have

$$\|\mathcal{O}\|_2\leq\|\Psi\Psi^{\intercal}/p\|_2\|H\|_2\|\widetilde{H}\|_2\frac{1}{n}\|\widehat{\Lambda}^{-2}\|_2.$$

With probability larger than $1-p^{-c}-\exp(-cn)$,

$$\|\widehat{\Lambda}^{-2}\|_2\lesssim\frac{p}{\lambda_q^2(\Psi)} \quad\text{and}\quad \|H\|_2\lesssim\sqrt{n}.$$

Applying the above inequality together with the fact that $\|\Psi\Psi^{\intercal}/p\|_2\leq\lambda_1^2(\Psi)/p$, $\|\widetilde{H}\|_2=\sqrt{n}$ and $\lambda_1(\Psi)/\lambda_q(\Psi)\leq C$, we establish that, with probability larger than $1-p^{-c}-\exp(-cn)$,

$$\|\mathcal{O}\|_2\leq C',$$

for some positive constant $C'>0$. That is, $\mathbb{P}(\mathcal{G}_{10})\geq 1-p^{-c}-\exp(-cn)$.

Control of $\mathcal{G}_{11}\cap\mathcal{G}_{12}$. The proofs follows from the fact that $E_{t,j}E_{t,l}-(\Sigma_E)_{j,l}$ and $H_{t,j}E_{t,l}$ are zero mean sub-exponential random variable. We apply Corollary 5.17 of [57] and the union bound to establish $\mathbb{P}(\mathcal{G}_{11}\cap\mathcal{G}_{12})\geq 1-p^{-c}$ for some positive constant $c>0$.

## APPENDIX C: ADDITIONAL PROOFS

**C.1. Proof of Proposition 2.** We note that

$$\mathcal{Q}y - \mathcal{Q}X\widehat{\beta}^{init} = \mathcal{Q}e + \mathcal{Q}\Delta + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb,$$

where $\Delta_i = \psi^\intercal H_{i,\cdot} - b^\intercal X_{i,\cdot}$ for $1 \le i \le n$.

Then we have

$$\begin{aligned}\widehat{\sigma}_e^2 - \sigma_e^2 = \frac{\|\mathcal{Q}e\|_2^2}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2 + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\|\mathcal{Q}\Delta + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb\|_2^2 \\ + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal \mathcal{Q}^2\Delta + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal \mathcal{Q}^2 X(\beta - \widehat{\beta}^{init}) + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal \mathcal{Q}^2 Xb.\end{aligned} \tag{115}$$

The following analysis is to study the above decomposition term by term. First note that

$$\frac{\|\mathcal{Q}e\|_2^2}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2 = \frac{e^\intercal US^2U^\intercal e}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2.$$

By Lemma 11 in Section C.7, we establish that with probability larger than $1-\exp(-ct^2)$ for $0 < t \lesssim \mathrm{Tr}(\mathcal{Q}^4) \asymp n$,

$$\left|\frac{e^\intercal \mathcal{Q}e}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2\right| \lesssim t\frac{\sqrt{\mathrm{Tr}(\mathcal{Q}^4)}}{\mathrm{Tr}(\mathcal{Q}^2)}. \tag{116}$$

By (47), we show that

$$\mathbf{P}\left(\frac{1}{n}\|\Delta\|_2^2 \lesssim q\log p/p\right) \ge 1-(\log p)^{-1/2}. \tag{117}$$

Since $e_i$ is independent of $X_{i,\cdot}$ and $H_{i,\cdot}$, the term $\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2\Delta$ is of mean zero and variance

$$\frac{1}{\mathrm{Tr}^2(\mathcal{Q}^2)}\sigma_e^2\|\mathcal{Q}^2\Delta\|_2^2 \lesssim \frac{\sigma_e^2}{n^2}\|\Delta\|_2^2,$$

where the inequality follows from $\mathrm{Tr}(\mathcal{Q}^2) \asymp m \asymp n$ and$\|\mathcal{Q}^2\Delta\|_2 \le \|\Delta\|_2$. Together with (117), we establish that, with probability larger than $1-(\log p)^{-1/2} - \frac{1}{t^2}$ for some $t > 0$,

$$\left|\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2\Delta\right| \lesssim t\sqrt{\frac{q\log p}{np}}\sigma_e. \tag{118}$$

Since $\mathrm{Tr}(\mathcal{Q}^2) \asymp m \asymp n$ and$\|\mathcal{Q}\Delta\|_2 \le \|\Delta\|_2$, we have

(119)

$$\begin{aligned}\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\|\mathcal{Q}\Delta + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb\|_2^2 &\lesssim \frac{1}{n}\|\mathcal{Q}\Delta\|_2^2 + \frac{1}{n}\|\mathcal{Q}X(\beta-\widehat{\beta}^{init})\|_2^2 + \frac{1}{n}\|\mathcal{Q}Xb\|_2^2 \\ &\lesssim \frac{q\log p}{p} + M^2\frac{k\log p}{n} + \frac{1}{n}\|\mathcal{Q}Xb\|_2^2\end{aligned}$$

with probability larger than $1-(\log p)^{-1/2}$.

Recall that $\widetilde{W} \in \mathbb{R}^{p\times p}$ as a diagonal matrix with diagonal entries as $\widetilde{W}_{l,l} = \|\mathcal{Q}X_{\cdot,l}\|_2/\sqrt{n}$ for $1 \le l \le p$. We establish that

(120)

$$\left|\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2X(\beta-\widehat{\beta}^{init})\right| \lesssim \|\frac{1}{n}e^\intercal\mathcal{Q}^2X\widetilde{W}^{-1}\|_\infty\|\widetilde{W}(\beta-\widehat{\beta}^{init})\|_1 \lesssim M^2k\lambda^2 + \left(\frac{\|\mathcal{Q}Xb\|_2}{\sqrt{n}}\right)^2,$$

where the last inequality follows from (41) and (153).

Finally, we control $\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)} e^{\intercal} \mathcal{Q}^2 X b$, which has mean zero and variance

$$\mathbb{E}\left(\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)} e^{\intercal} \mathcal{Q}^2 X b\right)^2 \lesssim \frac{1}{n^2}\sigma_e^2 b^{\intercal} X^{\intercal} \mathcal{Q}^4 X b \leq \frac{\sigma_e^2}{n^2}\|X^{\intercal} \mathcal{Q}^2 X\|_2 \|b\|_2^2$$

and hence with probability larger than $1-\frac{1}{t^2}$ for some $t>0$,

$$\frac{1}{n} e^{\intercal} \mathcal{Q}^2 X b \lesssim \frac{t}{\sqrt{n}} \frac{1}{\sqrt{n}} \|\mathcal{Q} X\|_2 \|b\|_2. \tag{121}$$

A combination of the decomposition (115) and the error bounds (116), (118), (119), (120), (121) and (44) leads to Proposition 2.

**C.2. Proof of Proposition 3.** By the Wely's inequality, we have that, for $1 \leq l \leq m$,

$$|\lambda_l(X) - \lambda_l(H\Psi)| = |\lambda_l(H\Psi + E) - \lambda_l(H\Psi)| \leq \lambda_1(E). \tag{122}$$

By Theorem 5.39 and equation (5.26) in [57] and $\lambda_{\max}(\Sigma_E) \leq C_0$, with probability larger than $1-\exp(-cn)$ for some $c>0$,

$$\lambda_{\max}(E) \lesssim \max\{\sqrt{n}, \sqrt{p}\}.$$

Note that $\lambda_l\left(\frac{1}{n}XX^{\intercal}\right) = \frac{1}{n}\lambda_l^2(X)$. Since $\lambda_{q+1}(H\Psi)=0$, we establish the proposition by applying (122).

**C.3. Proof of Lemma 1.** We express the model (2) as

$$X_{1,j} = \Psi_j^{\intercal} H_{1,\cdot} + E_{1,j}, \quad X_{1,-j} = \Psi_{-j}^{\intercal} H_{1,\cdot} + E_{1,-j},$$

where $\Psi_j \in \mathbb{R}^q$ denotes the $j$-th column of $\Psi_j$ and $\Psi_{-j} \in \mathbb{R}^{q\times(p-1)}$ denotes the sub-matrix of $\Psi$ except for the $j$-th column. We define $B = \mathbb{E}E_{1,-j}E_{1,-j}^{\intercal}$. Since $\mathrm{Cov}(H_{i,\cdot}) = \mathrm{I}_{q\times q}$ and the components of $H_{i,\cdot}$ are uncorrelated with the components of $\mathbb{E}_{i,\cdot}$, then we have

$$\gamma = [\mathbb{E}(X_{1,-j}X_{1,-j}^{\intercal})]^{-1}\mathbb{E}(X_{1,-j}X_{1,j}) = \left(\Psi_{-j}^{\intercal}\Psi_{-j} + B\right)^{-1}\left(\Psi_{-j}^{\intercal}\Psi_j + \mathbb{E}E_{1,j}E_{1,-j}\right). \tag{123}$$

We apply Woodbury matrix identity and then have

$$\left(\Psi_{-j}^{\intercal}\Psi_{-j} + B\right)^{-1} = B^{-1} - B^{-1}\Psi_{-j}^{\intercal}\left(\mathrm{I} + \Psi_{-j}B^{-1}\Psi_{-j}^{\intercal}\right)^{-1}\Psi_{-j}B^{-1}. \tag{124}$$

We combine the above two equalities and establish the decomposition $\gamma = \gamma^E + \gamma^A$ with

$$\gamma^E = B^{-1}\mathbb{E}E_{1,j}E_{1,-j}$$

and

$$\gamma^A = \left(\Psi_{-j}^{\intercal}\Psi_{-j} + B\right)^{-1}\Psi_{-j}^{\intercal}\Psi_j - B^{-1}\Psi_{-j}^{\intercal}\left(\mathrm{I} + \Psi_{-j}B^{-1}\Psi_{-j}^{\intercal}\right)^{-1}\Psi_{-j}\gamma^E. \tag{125}$$

Proof of (32). We define $D = \Psi_{-j}B^{-\frac{1}{2}} \in \mathbb{R}^{q\times(p-1)}$ and hence the first component on the right hand side of (125) can be expressed as

$$\left(\Psi_{-j}^{\intercal}\Psi_{-j} + B\right)^{-1}\Psi_{-j}^{\intercal}\Psi_j = B^{-\frac{1}{2}}\left(D^{\intercal}D + \mathrm{I}\right)^{-1}D^{\intercal}\Psi_j.$$

By Woodbury matrix identity, we have

$$\left(D^{\intercal}D + \mathrm{I}\right)^{-1}D^{\intercal} = \left(\mathrm{I} - D^{\intercal}(\mathrm{I} + DD^{\intercal})^{-1}D\right)D^{\intercal} = D^{\intercal}(\mathrm{I} + DD^{\intercal})^{-1}$$

and hence

$$\left(\Psi_{-j}^{\intercal}\Psi_{-j}+B\right)^{-1}\Psi_{-j}^{\intercal}\Psi_{j}=B^{-\frac{1}{2}}D^{\intercal}(\mathrm{I}+DD^{\intercal})^{-1}\Psi_{j}. \tag{126}$$

The second component on the right hand side of (125) can be expressed as

$$B^{-\frac{1}{2}}D^{\intercal}\left(\mathrm{I}+DD^{\intercal}\right)^{-1}\Psi_{-j}\gamma^{E}.$$

Together with (126), we simplify (125) as

$$\gamma^{A}=B^{-\frac{1}{2}}D^{\intercal}\left(\mathrm{I}+DD^{\intercal}\right)^{-1}\left(\Psi_{j}+\Psi_{-j}\gamma^{E}\right). \tag{127}$$

Under the assumption that $c_0\leq\lambda_{\min}(\Omega_E)\leq\lambda_{\max}(\Omega_E)\leq C_0$, we introduce the SVD for D as $D=U(D)\Lambda(D)V(D)^{\intercal}$, where $U(D),\Lambda(D)\in\mathbb{R}^{q\times q}$ and $V(D)\in\mathbb{R}^{(p-1)\times q}$. Since

$$D^{\intercal}\left(\mathrm{I}+DD^{\intercal}\right)^{-1}=V(D)\Lambda(D)(\Lambda(D)^{2}+\mathrm{I})^{-1}U(D)^{\intercal},$$

it follows from (127) that

$$\|\gamma^{A}\|_{2}\leq\|B^{-\frac{1}{2}}\|_{2}\max_{1\leq l\leq q}\frac{|\lambda_{l}(D)|}{\lambda_{l}^{2}(D)+1}\|\Psi_{j}+\Psi_{-j}\gamma^{E}\|_{2}, \tag{128}$$

where $\lambda_l(D)$ is the $l$-th largest singular value of $D$ in absolute value. By the condition $c_0\leq\lambda_{\min}(\Omega_E)\leq\lambda_{\max}(\Omega_E)\leq C_0$, we have $\frac{1}{C_0}\mathrm{I}\preceq B=\mathbb{E}E_{1,-j}E_{1,-j}^{\intercal}\preceq\frac{1}{c_0}\mathrm{I}$. We further have $c_0\lambda_l^2(\Psi_{-j})\leq\lambda_l^2(D)\leq C_0\lambda_l^2(\Psi_{-j})$ for $1\leq l\leq q$ and establish the first inequality of (32). The second inequality of (32) follows from condition (A2).
Proof of (33) We fix $1\leq i\leq n$ and $1\leq j\leq p$. Recall that

$$\eta_{i,j}=X_{i,j}-X_{i,-j}^{\intercal}\gamma=\Psi_{j}^{\intercal}H_{i,\cdot}-(\Psi_{-j}^{\intercal}H_{i,\cdot})^{\intercal}\gamma+E_{i,j}-E_{i,-j}^{\intercal}\gamma^{E}-E_{i,-j}^{\intercal}\gamma^{A},$$

$$\nu_{i,j}=E_{i,j}-E_{i,-j}^{\intercal}\gamma^{E},$$

and

$$\delta_{i,j}=\eta_{i,j}-\nu_{i,j}=\Psi_{j}^{\intercal}H_{i,\cdot}-(\Psi_{-j}^{\intercal}H_{i,\cdot})^{\intercal}\gamma-E_{i,-j}^{\intercal}\gamma^{A}.$$

Since $E_{i,\cdot}$ is uncorrelated with $H_{i,\cdot}$ and $\nu_{i,j}$ is uncorrelated with $E_{i,-j}$ and $H_{i,\cdot}$, we have $\nu_{i,j}$ to be uncorrelated with $\delta_{i,j}$. Hence we have

$$\mathrm{Var}(\eta_{i,j})=\mathrm{Var}(\nu_{i,j})+\mathrm{Var}(\delta_{i,j}). \tag{129}$$

By the expression of $\gamma$ in (123), we express $\mathrm{Var}(\eta_{i,j})=\mathrm{Var}(X_{i,j})-\mathrm{Var}(X_{i,-j}^{\intercal}\gamma)$ as
(130)

$$\begin{aligned}
&\|\Psi_{j}\|_{2}^{2}+(\Sigma_{E})_{j,j}-\left(\Psi_{-j}^{\intercal}\Psi_{j}+\mathbb{E}E_{1,j}E_{1,-j}\right)^{\intercal}\left(\Psi_{-j}^{\intercal}\Psi_{-j}+B\right)^{-1}\left(\Psi_{-j}^{\intercal}\Psi_{j}+\mathbb{E}E_{1,j}E_{1,-j}\right)\\
=&\|\Psi_{j}\|_{2}^{2}+(\Sigma_{E})_{j,j}-\left(\Psi_{-j}^{\intercal}\Psi_{j}+\mathbb{E}E_{1,j}E_{1,-j}\right)^{\intercal}B^{-1}\left(\Psi_{-j}^{\intercal}\Psi_{j}+\mathbb{E}E_{1,j}E_{1,-j}\right)\\
+&\left(\Psi_{-j}^{\intercal}\Psi_{j}+\mathbb{E}E_{1,j}E_{1,-j}\right)^{\intercal}B^{-1}\Psi_{-j}^{\intercal}\left(\mathrm{I}+\Psi_{-j}B^{-1}\Psi_{-j}^{\intercal}\right)^{-1}\Psi_{-j}B^{-1}\left(\Psi_{-j}^{\intercal}\Psi_{j}+\mathbb{E}E_{1,j}E_{1,-j}\right).
\end{aligned}$$

where the equation follows from (124). Note that

$$\mathrm{Var}(\nu_{i,j})=(\Sigma_{E})_{j,j}-(\mathbb{E}E_{1,j}E_{1,-j})^{\intercal}B^{-1}(\mathbb{E}E_{1,j}E_{1,-j}).$$

Together with (129) and (130), we obtain

(131)

$$\begin{aligned}
&\mathrm{Var}(\delta_{i,j}) = \|\Psi_j\|_2^2 - \Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \Psi_j - 2\Psi_j^\intercal \Psi_{-j} \gamma^E \\
&+ \left(\Psi_{-j}^\intercal \Psi_j + \mathbb{E} E_{1,j} E_{1,-j}\right)^\intercal B^{-1} \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} B^{-1} \left(\Psi_{-j}^\intercal \Psi_j + \mathbb{E} E_{1,j} E_{1,-j}\right) \\
&= \|\Psi_j\|_2^2 - \Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \Psi_j - 2\Psi_j^\intercal \Psi_{-j} \gamma^E \\
&+ \Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \Psi_j \\
&+ 2\Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E + (\gamma^E)^\intercal \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E.
\end{aligned}$$

Note that

$$\|\Psi_j\|_2^2 = \Psi_j^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right) \Psi_j.$$

We have

$$\begin{aligned}
&\|\Psi_j\|_2^2 + \Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \Psi_j \\
&= \Psi_j^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_j + \Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \Psi_j.
\end{aligned} \tag{132}$$

Note that

$$\begin{aligned}
&\Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E - \Psi_j^\intercal \Psi_{-j} \gamma^E \\
=&\Psi_j^\intercal \Psi_{-j} B^{-1} \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E \\
&- \Psi_j^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right) \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E \\
=&- \Psi_j^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E.
\end{aligned}$$

Together with (131) and (132), we establish

$$\begin{aligned}
\mathrm{Var}(\delta_{i,j}) &= \Psi_j^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_j + (\gamma^E)^\intercal \Psi_{-j}^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E \\
&- 2\Psi_j^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} \Psi_{-j} \gamma^E \\
&= (\Psi_j - \Psi_{-j} \gamma^E)^\intercal \left(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal\right)^{-1} (\Psi_j - \Psi_{-j} \gamma^E).
\end{aligned}$$

We establish (33) by applying condition (A2) and the following inequality

$$\lambda_{\min}(\mathrm{I} + \Psi_{-j} B^{-1} \Psi_{-j}^\intercal) \geq 1 + C\lambda_q^2(\Psi_{-j}),$$

for some positive constant $C > 0$.

**C.4. Proof of Lemma 2.** The proof of this lemma is similar to Lemma 1 in terms of controlling $\|b\|_2$. We start with the exact expression of $b$

$$b = \Sigma_X^{-1} \Psi^\intercal \phi = (\Sigma_E + \Psi^\intercal \Psi)^{-1} \Psi^\intercal \phi.$$

By apply the Woodbury matrix inverse formula, we have

$$b = \Sigma_E^{-1} \Psi^\intercal \left(\mathrm{I} + \Psi \Sigma_E^{-1} \Psi^\intercal\right)^{-1} \phi.$$

We define $D_E = \Psi\Sigma_E^{-1/2} \in \mathbb{R}^{q\times p}$ and hence we have

$$b = \Sigma_E^{-1/2} D_E^{\intercal}(\mathrm{I} + D_E D_E^{\intercal})^{-1}\phi,$$

and

$$b_j = (\Omega_E)_{\cdot,j}^{\intercal}\Psi^{\intercal}(\mathrm{I} + D_E D_E^{\intercal})^{-1}\phi. \tag{133}$$

Hence, we control $\|b\|_2$ as

$$\|b\|_2 \leq \sqrt{C_0}\max_{1\leq l\leq q}\frac{\lambda_l(D_E)}{1+\lambda_l^2(D_E)}\|\phi\|_2 \lesssim \frac{\sqrt{q}(\log p)^{1/4}}{\lambda_q(\Psi)}.$$

where the last inequality follows from the fact $c_0\lambda_j^2(\Psi) \leq \lambda_j^2(D_E) \leq C_0\lambda_j^2(\Psi)$ and the condition (A2). Similarly, we apply condition (A2) and control $|b_j|$ as

$$|b_j| \leq \|\Psi(\Omega_E)_{\cdot,j}\|_2\frac{1}{1+\lambda_q^2(D_E)}\|\phi\|_2 \lesssim \frac{q\sqrt{\log p}}{\lambda_q^2(\Psi)}.$$

It follows from Woodbury matrix inverse formula that

$$\Psi\Sigma_X^{-1}\Psi^{\intercal} = \Psi^{\intercal}\Sigma_E^{-1}\Psi(\mathrm{I}_q + \Psi\Sigma_E^{-1}\Psi^{\intercal})^{-1}, \tag{134}$$

and hence

$$\sigma_\epsilon^2 - \sigma_e^2 = \phi^{\intercal}\left(\mathrm{I}_q - \Psi\Sigma_X^{-1}\Psi^{\intercal}\right)\phi = \phi^{\intercal}(\mathrm{I}_q + \Psi\Sigma_E^{-1}\Psi^{\intercal})^{-1}\phi.$$

We establish (35) by applying condition (A2) and the following inequality

$$\lambda_{\min}(\mathrm{I} + \Psi\Sigma_E^{-1}\Psi^{\intercal}) \geq 1 + C\lambda_q^2(\Psi),$$

for some positive constant $C > 0$.

**C.5. Proof of Proposition 4.** Define $W \in \mathbb{R}^{p\times p}$ as a diagonal matrix with diagonal entries as $W_{l,l} = \|\mathcal{P}^{(j)}X_{\cdot,l}\|_2/\sqrt{n}$ for $1 \leq l \leq p$. For the vector $a \in \mathbb{R}^{p-1}$, we define the weighted $\ell_1$ norm $\|a\|_{1,w} = \sum_{l\neq j}\frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}}|a_l| = \|(W_{-l,-l})a\|_1$. Define the event

$$\mathcal{A}_0 = \left\{c \leq \frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}} \leq CM \quad \text{for } 1 \leq l \leq p\right\}, \tag{135}$$

for some positive constants $C > 0$ and $c > 0$. On the event $\mathcal{A}_0$, we have

$$c\|a\|_1 \leq \|a\|_{1,w} \leq CM\|a\|_1. \tag{136}$$

We now show that $\mathbb{P}(\mathcal{A}_0) \geq 1 - p^{-c} - \exp(-cn)$, for some positive constant $c > 0$. By the construction of $\mathcal{P}^{(j)}$, we have

$$\frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}} \leq \frac{\|X_{\cdot,l}\|_2}{\sqrt{n}}.$$

Following from the fact that $X_{i,l}$ is of sub-Gaussian norm $M$, we apply the Corollary 5.17 in [57] and establish that, with probability larger than $1 - p^{-c} - \exp(-cn)$,

$$\frac{\|X_{\cdot,l}\|_2}{\sqrt{n}} \lesssim \sqrt{\mathrm{Var}(X_{1,l})}(1 + M\sqrt{\log p/n}) \lesssim M, \tag{137}$$

where the last inequality follows from the definition of sub-Gaussian norm and $M\sqrt{\log p/n} \leq C$ for some positive constant $C > 0$. It follows from condition (A4) that

$$\min_{l\neq j}\frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}} \geq \sqrt{\tau_*}. \tag{138}$$

Recall the definitions

$$\eta_j = (\eta_{1,j}, \dots, \eta_{n,j})^{\intercal} \in \mathbb{R}^n, \quad \nu_j = (\nu_{1,j}, \dots, \nu_{n,j})^{\intercal} \in \mathbb{R}^n \quad \text{and} \quad \delta_j = \eta_j - \nu_j.$$

In the following, we shall choose the tuning parameter $\lambda_0$ such that

$$\lambda_0 \geq \|\frac{1}{n}\eta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_{\infty}.$$

Since $\nu_{i,j} = E_{i,j} - (\gamma^E)^{\intercal} E_{i,-j}$ is sub-Gaussian and independent of $X_{i,-j}$, we apply Proposition 5.10 in [57] and the maximum inequality to establish

$$\mathbb{P}\left(\|\frac{1}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_{\infty} \geq A_0\sigma_j\sqrt{\log p/n}\right) \leq e \cdot p^{1-c(A_0/C_1)^2} \tag{139}$$

for some positive constants $A_0 > 0$ and $c > 0$. We then control $\|\frac{1}{n}\delta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_{\infty}$ by the inequality

$$\|\frac{1}{n}\delta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_{\infty} \leq \frac{1}{\sqrt{n}}\|\delta_j\|_2$$

and the upper bound for $\frac{1}{n}\mathbb{E}\|\delta_j\|_2^2$ in (33). As a consequence, we have

$$\mathbb{P}\left(\|\frac{1}{n}\delta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_{\infty} \geq \frac{1}{1+c}\sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}}\right) \lesssim (\log p)^{-1/2}$$

for any positive constant $c > 0$. Together with (139), we then choose

$$\lambda_0 = A_0\sigma_j\sqrt{\frac{\log p}{n}} + \frac{1}{1+c}\sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}} \quad \text{and} \quad \lambda_j \geq (1+c)\lambda_0,$$

and have

$$\mathbb{P}\left(\|\frac{1}{n}\eta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_{\infty} \leq \lambda_0\right) \geq 1 - C(\log p)^{-1/2}, \tag{140}$$

for some positive constant $C > 0$.

By the definition of the estimator $\widehat{\gamma}$, we have the following basic inequality,

$$\frac{1}{2n}\|\mathcal{P}^{(j)}(X_j - X_{-j}\widehat{\gamma})\|_2^2 + \lambda_j\|\widehat{\gamma}\|_{1,w} \leq \frac{1}{2n}\|\mathcal{P}^{(j)}\left(X_j - X_{-j}\gamma^E\right)\|_2^2 + \lambda_j\|\gamma^E\|_{1,w}. \tag{141}$$

By decomposing $X_j - X_{-j}\widehat{\gamma} = X_{-j}\gamma^A + \eta_j + X_{-j}\left(\gamma^E - \widehat{\gamma}\right)$, we simplify (141) as

$$\begin{aligned} &\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^E - \widehat{\gamma}\right)\|_2^2 + \lambda_j\|\widehat{\gamma}\|_{1,w} \leq \lambda_j\|\gamma^E\|_{1,w} \\ &\quad - \frac{1}{n}\eta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}\left(\gamma^E - \widehat{\gamma}\right) - \frac{1}{n}\left(\mathcal{P}^{(j)}X_{-j}\gamma^A\right)^{\intercal}\mathcal{P}^{(j)}X_{-j}\left(\gamma^E - \widehat{\gamma}\right). \end{aligned} \tag{142}$$

Regarding the right hand side of the above inequality, we apply (140) and establish that, with probability larger than $1 - C(\log p)^{-1/2}$ for some positive constant $C > 0$,

$$\begin{aligned} \left|\frac{1}{n}\eta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}\left(\gamma^E - \widehat{\gamma}\right)\right| &\leq \|\frac{1}{n}\eta_j^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_{\infty}\|W_{-j,-j}(\gamma^E - \widehat{\gamma})\|_1 \\ &\leq \lambda_0\|\gamma^E - \widehat{\gamma}\|_{1,w}. \end{aligned}$$

Additionally, we have

$$\left|\frac{1}{n}\left(\mathcal{P}^{(j)}X_{-j}\gamma^A\right)^{\intercal}\mathcal{P}^{(j)}X_{-j}\left(\gamma^E - \widehat{\gamma}\right)\right| \leq \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^E - \widehat{\gamma}\right)\|_2.$$

Then we further simply (142) as

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2^2+\lambda_j\|\widehat{\gamma}\|_{1,w}\leq\lambda_j\|\gamma^{E}\|_{1,w}+\lambda_0\|\gamma^{E}-\widehat{\gamma}\|_{1,w}$$
$$+\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2.$$

Let $\mathcal{T}_j$ denote the support of $\gamma^E$. By the fact that $\|\gamma^{E}_{\mathcal{T}_j}\|_{1,w}-\|\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\leq\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}$ and $\|\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}=\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}$, then we establish

$$\begin{aligned}&\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2^2+\left(\lambda_j-\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\\&\leq\left(\lambda_j+\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}+\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2.\end{aligned}\tag{143}$$

The following analysis is based on (143) and divided into two cases depending on the dominating term on the right hand side of (143).
**Case 1**: We consider

$$\left(\lambda_j+\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\geq\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2$$

and then simplify (143) as

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2^2+\left(\lambda_j-\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\leq2\left(\lambda_j+\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}.\tag{144}$$

It follows from (144) that

$$\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\leq\frac{\lambda_j+\lambda_0}{\lambda_j-\lambda_0}\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}.$$

By the choices of $\lambda_j$ and $\lambda_0$, on the event $\mathcal{A}_0$, we establish

$$\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_1\leq CM\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_1,$$

for some positive constant $C>0$. By the restricted eigenvalue condition (22), we have

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2^2\geq\frac{\tau_*}{2}\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2^2.$$

Together with (144) and (136), we have

$$\begin{aligned}\frac{\tau_*}{2}\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2^2&\leq2\left(\lambda_j+\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\\&\leq2CM\left(\lambda_j+\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_1\\&\leq2CM\sqrt{|\mathcal{T}_j|}\left(\lambda_j+\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2,\end{aligned}$$

which leads to

$$\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2\lesssim\frac{M}{\tau_*}\sqrt{|\mathcal{T}_j|}\left(\lambda_j+\lambda_0\right)\quad\text{and}\quad\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_1\lesssim\frac{M}{\tau_*}|\mathcal{T}_j|\left(\lambda_j+\lambda_0\right).$$

On the event $\mathcal{A}_0$, the above inequality implies that

$$\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_1\lesssim\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\lesssim\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\lesssim M\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_1\lesssim\frac{M^2}{\tau_*}|\mathcal{T}_j|\left(\lambda_j+\lambda_0\right).\tag{145}$$

Together with (144), (145) implies that

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2^2 \lesssim \frac{M^2}{\tau_*}|\mathcal{T}_j|\left(\lambda_j+\lambda_0\right)^2. \tag{146}$$

We apply the restricted eigenvalue condition (22) again to establish

$$\|\gamma^{E}-\widehat{\gamma}\|_2 \lesssim M\sqrt{|\mathcal{T}_j|}\left(\lambda_j+\lambda_0\right). \tag{147}$$

**Case 2**: We consider

$$\left(\lambda_j+\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w} \leq \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2,$$

and then simplify (143) as

$$\begin{aligned}\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2^2+\left(\lambda_j-\lambda_0\right)\|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\\ \leq \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2.\end{aligned}$$

Then we derive

$$\frac{1}{\sqrt{n}}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^{E}-\widehat{\gamma}\right)\|_2 \lesssim \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2, \tag{148}$$

$$\|\gamma^{E}_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w} \lesssim \frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2^2}{\lambda_j+\lambda_0} \quad \text{and} \quad \|\gamma^{E}_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w} \lesssim \frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2^2}{\lambda_j-\lambda_0}. \tag{149}$$

Then, on the event $\mathcal{A}_0$, we have

$$\|\gamma^{E}-\widehat{\gamma}\|_2 \leq \|\gamma^{E}-\widehat{\gamma}\|_1 \lesssim \|\gamma^{E}-\widehat{\gamma}\|_{1,w} \lesssim \frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2^2}{\lambda_j+\lambda_0}+\frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^{A}\|_2^2}{\lambda_j-\lambda_0}. \tag{150}$$

Finally, we establish (36) by combining (145) and (149); establish (37) by combining (147) and (150); establish (38) by combining (146) and (148);

### C.6. Proof of Proposition 5.

The proof of Proposition 5 is similar to the proof of Proposition 4 in Section C.5. In the following, we prove Proposition 5 and mainly highlight its difference from the proof of Proposition 4 in Section C.5.

Define $\widetilde{W} \in \mathbb{R}^{p\times p}$ as a diagonal matrix with diagonal entries as $\widetilde{W}_{l,l}=\|\mathcal{Q}X_{\cdot,l}\|_2/\sqrt{n}$ for $1\leq l\leq p$.With a slight abuse of notation, for $a\in\mathbb{R}^p$, we define $\|a\|_{1,w}=\sum_{l=1}^{p}\frac{\|\mathcal{Q}X_{\cdot,l}\|_2}{\sqrt{n}}|a_l|$. Define the event

$$\mathcal{A}_1=\left\{c\leq\frac{\|\mathcal{Q}X_{\cdot,l}\|_2}{\sqrt{n}}\leq CM \quad \text{for } 1\leq l\leq p\right\},$$

for some positive constants $C>c>0$. On the event $\mathcal{A}_1$, we have (136). Similar to the control of $\mathcal{A}_0$ defined in (135), we can show that $\mathbb{P}(\mathcal{A}_1)\geq 1-p^{-c}-\exp(-cn)$ for some positive constant $c>0$.

The main part of the proof is to calculate the tuning parameter $\lambda$ such that

$$\lambda\geq(1+c)\|\frac{1}{n}\epsilon^{\intercal}\mathcal{Q}^2X\widetilde{W}^{-1}\|_\infty$$

for a small positive constant $c>0$. Note that $\epsilon=e+\Delta$ with $\Delta_i=\psi^{\intercal}H_{i,\cdot}-b^{\intercal}X_{i,\cdot}$. Since $e_i$ is independent of $X_{i,\cdot}$, we apply Proposition 5.10 in [57] and the maximum inequality to establish

$$\mathbb{P}\left(\|\frac{1}{n}e^{\intercal}\mathcal{Q}^2X\widetilde{W}^{-1}\|_\infty\geq A_0\sigma_e\sqrt{\log p/n}\right)\leq e\cdot p^{-c(A_0/C_1)^2}, \tag{151}$$

for some positive constants $c>0$ and $A_0>0$. We then control the other part $\|\frac{1}{n}\Delta^{\intercal}\mathcal{Q}^2 X\widetilde{W}^{-1}\|_{\infty}$ by the inequality

$$\|\frac{1}{n}\Delta^{\intercal}\mathcal{Q}^2 X\widetilde{W}^{-1}\|_{\infty}\leq \frac{1}{\sqrt{n}}\|\Delta\|_2$$

and the upper bound for $\frac{1}{n}\mathbb{E}\|\Delta\|_2^2$ in (47). As a consequence, we have

$$\mathbb{P}\left(\|\frac{1}{n}\Delta^{\intercal}\mathcal{Q}^2 X\widetilde{W}^{-1}\|_{\infty}\geq \frac{1}{1+c}\sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi)}}\right)\lesssim (\log p)^{-1/2}, \tag{152}$$

for any positive constant $c>0$. We then choose

$$\lambda\geq A\sigma_e\sqrt{\frac{\log p}{n}}+\sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi)}}\quad\text{with}\quad A=(1+c)A_0.$$

We combine (151) and (152) and establish that

$$\mathbb{P}\left((1+c_0)\|\frac{1}{n}\epsilon^{\intercal}\mathcal{Q}^2 XW^{-1}\|_{\infty}\leq\lambda\right)\geq 1-C(\log p)^{-1/2}-p^{-c}, \tag{153}$$

for some positive constant $C>0$.

By the definition of $\widehat{\beta}^{init}$, we establish the basic inequality in a similar fashion to (141)

$$\frac{1}{2n}\|\mathcal{Q}(Y-X\widehat{\beta}^{init})\|_2^2+\lambda\|\widehat{\beta}^{init}\|_{1,w}\leq\frac{1}{2n}\|\mathcal{Q}\left(Y-X\beta\right)\|_2^2+\lambda\|\beta\|_{1,w}. \tag{154}$$

We can apply the similar argument from (141) to (150) by replacing $\mathcal{P}^{(j)}$, $X_j$, $X_{-j}$, $\widehat{\gamma}$, $\gamma^E$, $\gamma^A$ with $\mathcal{Q}$, $Y$, $X$, $\widehat{\beta}^{init}$, $\beta$, $b$, respectively. We replace the tuning parameters $\lambda_j$ and $\lambda_0$ by $\lambda$ and $\frac{1}{1+c_0}\lambda$, respectively. Then we establish Proposition 5.

### C.7. Proof of Lemma 3.

We introduce the following lemma about the concentration of quadratic forms, which is Theorem 1.1 in [52].

LEMMA 11. (Hanson-Wright inequality) *Let $\xi\in\mathbb{R}^n$ be a random vector with independent sub-Gaussian components $\xi_i$ with zero mean and sub-Gaussian norm $K$. Let $A$ be an $n\times n$ matrix. Then for every $t\geq 0$,*

$$\mathbf{P}\left(|\xi^{\intercal}A\xi-\mathbb{E}\xi^{\intercal}A\xi|>t\right)\leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2},\frac{t}{K^2\|A\|_2}\right)\right]. \tag{155}$$

For the high-dimensional setting where $p/n\to c^*\in(0,\infty]$, we have $m\asymp n$ for $m=\min\{n,p-1\}$. We also note $\mathrm{Tr}[(\mathcal{P}^{(j)})^l]\asymp m$ for $l=2,4,8$.

C.7.1. *Proof of* (50). We decompose $\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_j$ as

$$\begin{aligned}\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_j&=\frac{1}{n}(\mathcal{P}^{(j)}\eta_j)^{\intercal}\mathcal{P}^{(j)}\eta_j+\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}\gamma\\&-\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E))^{\intercal}\mathcal{P}^{(j)}\eta_j+\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^{\intercal}\mathcal{P}^{(j)}\eta_j,\end{aligned} \tag{156}$$

where $\eta_j=(\eta_{1,j},\ldots,\eta_{n,j})^{\intercal}\in\mathbb{R}^n$.

In the following, we control the right hand side of (156) term by term. Since $\eta_j=\nu_j+\delta_j$, we have

$$\frac{1}{n}(\mathcal{P}^{(j)}\eta_j)^{\intercal}\mathcal{P}^{(j)}\eta_j=\frac{1}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^2\nu_j+\frac{2}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^2\delta_j+\frac{1}{n}\delta_j^{\intercal}(\mathcal{P}^{(j)})^2\delta_j.$$

By applying (155) with $A=(\mathcal{P}^{(j)})^2$, then with probability larger than $1-2\exp(-ct^2)$ for $0<t\lesssim \mathrm{Tr}[(\mathcal{P}^{(j)})^4]\asymp n$,

$$\left|\frac{1}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^2\nu_j-\mathrm{Tr}[(\mathcal{P}^{(j)})^2]\cdot\frac{\sigma_j^2}{n}\right|\lesssim t\frac{\sqrt{\mathrm{Tr}[(\mathcal{P}^{(j)})^4]}}{n}\lesssim t\frac{\sqrt{m}}{n}. \tag{157}$$

Since $\left|\delta_j^{\intercal}(\mathcal{P}^{(j)})^2\delta_j\right|\leq\|\delta_j\|_2^2$, we apply the upper bound (33) for $\frac{1}{n}\mathbb{E}\|\delta_j\|_2^2$ and the Markov inequality to establish

$$\mathbb{P}\left(\frac{1}{n}\|\delta_j\|_2^2\gtrsim\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}\right)\leq(\log p)^{-1/2}. \tag{158}$$

Hence, we have, with probability larger than $1-2\exp(-ct^2)-(\log p)^{-1/2}$,

$$\begin{aligned}\left|\frac{2}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^2\delta_j\right|&\leq 2\sqrt{\frac{1}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^2\nu_j}\sqrt{\frac{1}{n}\delta_j^{\intercal}(\mathcal{P}^{(j)})^2\delta_j}\\&\lesssim\sqrt{\left(\mathrm{Tr}[(\mathcal{P}^{(j)})^2]\cdot\frac{\sigma_j^2}{n}+Ct\frac{\sqrt{m}}{n}\right)\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}},\end{aligned}$$

for some positive constant $C>0$. Combined with (157) and (158), we apply the fact that $\mathrm{Tr}[(\mathcal{P}^{(j)})^2]\asymp n$ and establish that, with probability larger than $1-2\exp(-ct^2)-(\log p)^{-1/2}$ for $0<t\lesssim n$,

$$\left|\frac{1}{n}\eta_j^{\intercal}(\mathcal{P}^{(j)})^2\eta_j-\mathrm{Tr}[(\mathcal{P}^{(j)})^2]\cdot\frac{\sigma_j^2}{n}\right|\lesssim t\frac{\sqrt{m}}{n}+\sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}}. \tag{159}$$

By the KKT condition of (9), we establish

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}\gamma\right|\leq\|\gamma\|_1\|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}\|_\infty\leq\lambda_j\max_{l\neq j}\frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}}\|\gamma\|_1,$$

where $\lambda_j$ is defined in (39). We control the right hand side as

$$\left(\|\gamma^E\|_1+\|\gamma^A\|_1\right)\lambda_j\leq\sqrt{s}\|\gamma^E\|_2\lambda_j+\sqrt{p}\|\gamma^A\|_2\lambda_j.$$

On the event $\mathcal{A}_0$ defined in (135), we obtain

$$\begin{aligned}\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}\gamma\right|&\leq M\cdot(\sqrt{s}\|\gamma^E\|_2\lambda_j+\sqrt{p}\|\gamma^A\|_2\lambda_j)\\&\lesssim M\cdot\left(\sqrt{s}\|\gamma^E\|_2\lambda_j+\sqrt{\frac{p\log p}{n}+\frac{qp\log p}{\lambda_q^2(\Psi_{-j})}}\cdot\frac{\sqrt{q}(\log p)^{1/4}}{\lambda_q(\Psi_{-j})}\right).\end{aligned} \tag{160}$$

where the last bound follows from the definition of $\lambda_j$ in (39) and the upper bound for $\|\gamma^A\|_2$ in (32). We apply Hölder's inequality and establish

$$\begin{aligned}\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E))^{\intercal}\mathcal{P}^{(j)}\eta_j\right|&\leq\|W_{-j,-j}(\widehat{\gamma}-\gamma^E)\|_1\|\frac{1}{n}\eta_j^{\intercal}(\mathcal{P}^{(j)})^2X_{-j}(W_{-j,-j})^{-1}\|_\infty\\&\lesssim\frac{M^2}{\tau_*}s\lambda_j^2+\frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2}{n}\\&\lesssim\frac{M^2}{\tau_*}s\lambda_j^2+\frac{q\sqrt{\log p}}{1+\lambda_q^2(\Psi_{-j})}\cdot\max\left\{1,\frac{p}{n}\right\}\end{aligned} \tag{161}$$

where the second inequality follows from (36) and (140) and the last inequality follows from (40). Since $\nu_j$ is independent of $X_{-j}$, we show that $\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^\intercal \mathcal{P}^{(j)}\nu_j$ has mean zero and variance

(162)
$$\frac{\sigma_j^2}{n^2}(\gamma^A)^\intercal X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^\intercal \gamma^A \lesssim \frac{\|\gamma^A\|_2^2}{n}\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^\intercal\|_2 \lesssim \max\left\{1, \frac{p}{n}\right\} \cdot \frac{q\sqrt{\log p}}{n\lambda_q^2(\Psi_{-j})},$$

where the last inequality follows from the upper bound for $\|\gamma^A\|_2$ in (32) together with the property (P1). Then with probability larger than $1 - \frac{1}{t^2}$,

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^\intercal \mathcal{P}^{(j)}\nu_j\right| \lesssim \frac{t}{\sqrt{n}} \cdot \sqrt{\max\left\{1, \frac{p}{n}\right\} \cdot \frac{q(\log p)^{1/2}}{\lambda_q^2(\Psi_{-j})}}. \tag{163}$$

Note that, with probability larger than $1 - (\log p)^{-1/2}$,

$$\begin{aligned}\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^\intercal \mathcal{P}^{(j)}\delta_j\right| &\le \|\frac{1}{n}(\mathcal{P}^{(j)})^2 X_{-j}\gamma^A\|_2\|\delta_j\|_2 \\ &\lesssim \sqrt{\max\left\{1, \frac{p}{n}\right\} \cdot \frac{q\sqrt{\log p}}{n\lambda_q^2(\Psi_{-j})}} \cdot \sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}}\end{aligned} \tag{164}$$

where the last inequality follows from (162) and (158).

By (156), we combine the fact that $\mathrm{Tr}[(\mathcal{P}^{(j)})^2] \cdot \frac{\sigma_j^2}{n}$ is of a constant order and the upper bounds (159), (160), (161), (163) and (164). We establish (50) under the conditions $s\lambda_j^2 M^2 \to 0$ and

$$\lambda_q(\Psi_{-j}) \gg \max\left\{(1+M) \cdot \sqrt{\frac{qp}{n}}(\log p)^{3/4}, \sqrt{q(1+M)}p^{1/4}(\log p)^{3/8}\right\}.$$

Note that the above conditions are implied by (19) and $s \ll n/[M^2 \log p]$.

C.7.2. *Proof of* (51). Note that

$$\begin{aligned}\frac{1}{n}Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j = &\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^4\eta_j + 2\eta_j^\intercal(\mathcal{P}^{(j)})^4 X_{-j}(\gamma^E - \widehat{\gamma} + \gamma^A) \\ &+ \frac{1}{n}\|(\mathcal{P}^{(j)})^2 X_{-j}(\gamma^E - \widehat{\gamma} + \gamma^A)\|_2^2.\end{aligned} \tag{165}$$

By applying (155) with $A = (\mathcal{P}^{(j)})^4$, then with probability larger than $1 - 2\exp(-ct^2)$ for $0 < t \lesssim \mathrm{Tr}[(\mathcal{P}^{(j)})^8] \asymp n$,

$$\left|\frac{1}{n}\nu_j^\intercal(\mathcal{P}^{(j)})^4\nu_j - \mathrm{Tr}[(\mathcal{P}^{(j)})^4] \cdot \frac{\sigma_j^2}{n}\right| \lesssim t\frac{\sqrt{\mathrm{Tr}[(\mathcal{P}^{(j)})^8]}}{n} \lesssim t\frac{\sqrt{m}}{n}.$$

By a similar argument as in (159), we establish that, with probability larger than $1 - 2\exp(-ct^2) - (\log p)^{-1/2}$ for $0 < t \lesssim n$,

$$\left|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^4\eta_j - \mathrm{Tr}[(\mathcal{P}^{(j)})^4] \cdot \frac{\sigma_j^2}{n}\right| \lesssim t\frac{\sqrt{m}}{n} + \sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}}. \tag{166}$$

By a similar argument as (161), we have

$$\left|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^4 X_{-j}(\widehat{\gamma} - \gamma^E)\right| \lesssim \frac{M^2}{\tau_*}s\lambda_j^2 + \frac{q\sqrt{\log p}}{1+\lambda_q^2(\Psi_{-j})} \cdot \max\left\{1, \frac{p}{n}\right\}. \tag{167}$$

In addition, $\frac{1}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^4 X_{-j}\gamma^A$ has mean zero and variance

$$\frac{\sigma_j^2}{n^2}(\gamma^A)^{\intercal} X_{-j}(\mathcal{P}^{(j)})^8 X_{-j}^{\intercal}\gamma^A \lesssim \frac{\|\gamma^A\|_2^2}{n}\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^8 X_{-j}^{\intercal}\|_2,$$

and hence with probability larger than $1-\frac{1}{t^2}$ for any $t>0$,

$$\left|\frac{1}{n}\nu_j^{\intercal}(\mathcal{P}^{(j)})^4 X_{-j}\gamma^A\right| \lesssim \frac{t\|\gamma^A\|_2}{\sqrt{n}}\sqrt{\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^{\intercal}\|_2} \lesssim t\sqrt{\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{n\lambda_q^2(\Psi_{-j})}}, \tag{168}$$

where the last inequality follows from the upper bound for $\|\gamma^A\|_2$ in (32) together with the property (P1). Note that, with probability larger than $1-(\log p)^{-1/2}$,

$$\begin{aligned}\left|\frac{1}{n}\delta_j^{\intercal}(\mathcal{P}^{(j)})^4 X_{-j}\gamma^A\right| &\le \|\frac{1}{n}(\mathcal{P}^{(j)})^4 X_{-j}\gamma^A\|_2\|\delta_j\|_2\\ &\lesssim \sqrt{\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{n\lambda_q^2(\Psi_{-j})}}\cdot\sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}}\end{aligned} \tag{169}$$

where the last inequality follows from (162) and (158).

Note that

$$\begin{aligned}&\frac{1}{n}\|(\mathcal{P}^{(j)})^2 X_{-j}(\widehat{\gamma}-\gamma^E-\gamma^A)\|_2^2 \le \frac{1}{n}\|\mathcal{P}^{(j)} X_{-j}(\widehat{\gamma}-\gamma^E-\gamma^A)\|_2^2\\ &\lesssim \frac{1}{n}\|\mathcal{P}^{(j)} X_{-j}(\widehat{\gamma}-\gamma^E)\|_2^2+\frac{1}{n}\|\mathcal{P}^{(j)} X_{-j}\gamma^A\|_2^2.\end{aligned}$$

By applying (38) and (40), we establish that, with probability larger than $1-e\cdot p^{1-c(A_0/C_1)^2}-\exp(-cn)-(\log p)^{-1/2}$ for some positive constant $c>0$,

$$\frac{1}{n}\|(\mathcal{P}^{(j)})^2 X_{-j}(\widehat{\gamma}-\gamma^E-\gamma^A)\|_2^2 \lesssim \left(\frac{M}{\tau_*}\sqrt{s}\lambda_j+\sqrt{\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{n\lambda_q^2(\Psi_{-j})}}\right)^2. \tag{170}$$

By (165), we combine the fact that $\mathrm{Tr}[(\mathcal{P}^{(j)})^2]\cdot\frac{\sigma_j^4}{n}$ is of a constant order and the upper bounds (166), (167), (168), (169) and (170). We establish (51) under the condition

$$\lambda_q(\Psi_{-j}) \gg \sqrt{q}(\log p)^{1/4}\max\left\{\sqrt{\frac{p}{n}},(\log p)^{1/4}\right\} \quad \text{and} \quad s\lambda_j^2M^2\to 0.$$

Note that the above condition is implied by (19) and $s\ll n/[M^2\log p]$.

C.7.3. *Proof of* (52). Note that

$$|B_\beta| = \left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})}{(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_j}\right| \le \frac{\left|(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})\right|}{\sqrt{\sigma_e^2\cdot Z_j^{\intercal}(\mathcal{P}^{(j)})^4 Z_j}}.$$

It follows from Hölder's inequality and also the KKT condition of (9) that

$$\begin{aligned}\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})\right| &\le \|\beta_{-j}-\widehat{\beta}_{-j}^{init}\|_1\|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}\|_\infty\\ &\le \lambda_j\max_{l\neq j}\frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}}\|\beta_{-j}-\widehat{\beta}_{-j}^{init}\|_1.\end{aligned}$$

By the definition of the event $\mathcal{A}_0$ in (135) and the upper bound for $\|\widehat{\beta}^{init}-\beta\|_1$ in (41) and (44), with probability larger than $1-e\cdot p^{1-c(A/C_1)^2}-\exp(-cn)-(\log p)^{-1/2}$ for some positive constant $c>0$,

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}^{init}_{-j})\right|\lesssim M\left(\frac{M^2}{\tau_*}k\lambda_j\lambda+\frac{\lambda_j}{\lambda}\frac{q\sqrt{\log p}}{1+\lambda_q^2(\Psi)}\right).$$

Together with (51), we establish $B_\beta\xrightarrow{p}0$ under the condition

$$\lambda_q(\Psi)\gg[qM]^{1/2}(n\log p)^{1/4}\quad\text{and}\quad\sqrt{n}k\lambda_j\lambda[M]^3\to 0. \tag{171}$$

Note that the above condition is implied by (19) and $k\ll\sqrt{n}/[M^3\log p]$.

Now we control the other bias component

$$|B_b|=\left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}}{(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_j}+\frac{1}{\sqrt{V}}b_j\right|\leq\frac{\left|(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}\right|}{\sqrt{\sigma_e^2\cdot Z_j^{\intercal}(\mathcal{P}^{(j)})^4Z_j}}+\left|\frac{1}{\sqrt{V}}b_j\right|$$

We investigate $\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}$:

$$\begin{aligned}\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}&=\frac{1}{n}(\mathcal{P}^{(j)}\nu_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}+\frac{1}{n}(\mathcal{P}^{(j)}\delta_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}\\&\quad+\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E-\gamma^A))^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}.\end{aligned}\tag{172}$$

Note that $\frac{1}{n}(\mathcal{P}^{(j)}\nu_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}$ has mean zero and variance

$$\frac{\sigma_j^2}{n^2}(b_{-j})^{\intercal}X_{-j}(\mathcal{P}^{(j)})^4X_{-j}^{\intercal}b_{-j}\lesssim\frac{1}{n}\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4X_{-j}^{\intercal}\|_2\|b_{-j}\|_2^2\lesssim\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{n\lambda_q^2(\Psi)},$$

where the last inequality follows from the upper bound for $\|b_{-j}\|_2$ in (34) together with the property (P1). Hence with probability larger than $1-\frac{1}{t^2}$ for some $t>0$,

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}\nu_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}\right|\lesssim t\sqrt{\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{n\lambda_q^2(\Psi)}}.\tag{173}$$

Note that, with probability larger than $1-(\log p)^{-1/2}$,

$$\begin{aligned}\left|\frac{1}{n}(\mathcal{P}^{(j)}\delta_j)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}\right|&\leq\|\frac{1}{n}(\mathcal{P}^{(j)})^2X_{-j}\|_2\|b_{-j}\|_2\|\delta_j\|_2\\&\lesssim\sqrt{\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{\lambda_q^2(\Psi)}}\cdot\sqrt{\frac{q\log p}{1+\lambda_q^2(\Psi_{-j})}},\end{aligned}\tag{174}$$

where the last inequality follows from the upper bound for $\|\delta_j\|_2$ in (158), the upper bound for $\|b_{-j}\|_2$ in (34) together with the property (P1). In addition, we note the following two inequalities

$$\begin{aligned}\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^{\intercal}\mathcal{P}^{(j)}X_{-j}b_{-j}\right|&\leq\|\gamma^A\|_2\|b_{-j}\|_2\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^2X_{-j}^{\intercal}\|_2\\&\lesssim\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{\lambda_q(\Psi)\cdot\lambda_q(\Psi_{-j})},\end{aligned}\tag{175}$$

where the last inequality follows from the upper bound for $\|\gamma^A\|_2$ in (32), the upper bound for $\|b_{-j}\|_2$ in (34) together with the property (P1).

Note that

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E))^\intercal \mathcal{P}^{(j)}X_{-j}b_{-j}\right| \leq \frac{1}{\sqrt{n}}\|\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E)\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\|_2\|b_{-j}\|_2.$$

Furthermore, we apply the upper bound (38) and establish that, with probability larger than $1-e\cdot p^{1-c(A_0/C_1)^2}-\exp(-cn)-(\log p)^{-1/2}$ for some positive constant $c>0$,

$$\begin{aligned}
&\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E))^\intercal \mathcal{P}^{(j)}X_{-j}b_{-j}\right| \\
&\lesssim \left(\frac{M}{\tau_*}\sqrt{s}\lambda_j+\frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2}{\sqrt{n}}\right)\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\|_2\|b_{-j}\|_2 \qquad (176)\\
&\lesssim \left(\frac{M}{\tau_*}\sqrt{s}\lambda_j+\sqrt{\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{\lambda_q^2(\Psi_{-j})}}\right)\sqrt{\max\left\{1,\frac{p}{n}\right\}\cdot\frac{q\sqrt{\log p}}{\lambda_q^2(\Psi_{-j})}}.
\end{aligned}$$

where the last inequality follows from the upper bound for $\|\gamma^A\|_2$ in (32), the upper bound for $\|b_{-j}\|_2$ in (34) together with the property (P1).

By (32), (50) and (51), we have $\frac{|b_j|}{\sqrt{V}}\xrightarrow{p}0$ if $\sqrt{n}\frac{q\sqrt{\log p}}{1+\lambda_q^2(\Psi)}\to 0$.

We now combine the decomposition (172) and the upper bounds (173), (174), (175) and (176). Together with (50) and (51), we establish $B_b\xrightarrow{p}0$ under the condition

$$\lambda_q(\Psi)\geq\lambda_q(\Psi_{-j})\gg\sqrt{q}(\log p)^{1/4}\max\left\{\sqrt{\frac{p}{n}},n^{1/4}\right\}\quad\text{and}\quad\sqrt{s}\lambda_j M\to 0.$$

Note that the above condition is implied by (19) and $s\ll n/[M^2\log p]$.

**C.8. Proof of Lemma 4.** We first control the lower bound of $\lambda_q(\Psi)$ and the argument for $\lambda_q(\Psi_{-j})$ is similar. Note that $\lambda_q^2(\Psi)$ is the smallest eigenvalue of $\Psi\Psi^\intercal=\sum_{l=1}^p\Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal$. Since $\Psi_{\cdot,l}\in\mathbb{R}^q$ for $1\leq j\leq p$ are i.i.d. sub-Gaussian random vectors, it follows from (5.26) in [57], with probability larger than $1-p^{-c}$,

$$\|\frac{1}{p}\sum_{l=1}^p\Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal-\Sigma_\Psi\|_2\leq C\lambda_{\max}(\Sigma_\Psi)\sqrt{\frac{q+\log p}{p}},$$

for some positive constants $c,C>0$. This gives us that, with probability larger than $1-p^{-c}$,

$$\lambda_q^2(\Psi)=\lambda_{\min}(\sum_{j=1}^p\Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal)\gtrsim p\left(\lambda_{\min}(\Sigma_\Psi)-\lambda_{\max}(\Sigma_\Psi)\sqrt{\frac{q+\log p}{p}}\right). \qquad (177)$$

Similarly, we establish that, with probability larger than $1-p^{-c}$,

$$\lambda_q^2(\Psi_{-j})=\lambda_{\min}(\sum_{l\neq j}^p\Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal)\gtrsim(p-1)\left(\lambda_{\min}(\Sigma_\Psi)-\lambda_{\max}(\Sigma_\Psi)\sqrt{\frac{q+\log p}{p}}\right). \qquad (178)$$

In the following, we control $\Psi a$ for $a\in\mathbb{R}^p$ by noting that $\mathbb{E}\|\Psi a\|_2^2=\mathrm{Tr}(\Sigma_\Psi)\|a\|_2^2$. Hence, with probability larger than $1-\frac{1}{t^2}$, we have

$$\|\Psi a\|_2^2\leq t^2\mathrm{Tr}(\Sigma_\Psi)\|a\|_2^2\leq t^2q\lambda_{\max}(\Sigma_\Psi)\|a\|_2^2. \qquad (179)$$

By taking $a \in \mathbb{R}^p$ as $((\Omega_E)_{1,j}, \ldots, (\Omega_E)_{j-1,j}, 0, (\Omega_E)_{j+1,j}, \ldots, (\Omega_E)_{p,j})$, $e_j$ and $(\Omega_E)_{j,\cdot}$, we establish that with probability larger than $1 - \frac{1}{t^2}$,

$$\|\Psi_{-j}(\Omega_E)_{-j,j}\|_2 \lesssim t\sqrt{q}\sqrt{\lambda_{\max}(\Sigma_\Psi)}\|(\Omega_E)_{-j,j}\|_2 \tag{180}$$

$$\|\Psi_j\|_2 \lesssim t\sqrt{q}\sqrt{\lambda_{\max}(\Sigma_\Psi)} \tag{181}$$

$$\|\Psi(\Omega_E)_{\cdot,j}\|_2 \lesssim t\sqrt{q}\sqrt{\lambda_{\max}(\Sigma_\Psi)}\|(\Omega_E)_{\cdot,j}\|_2 \tag{182}$$

The lemma follows from a combination of (177), (178), (180), (181) and (182).

**C.9. Proof of Lemma 5.** The proof is a generalization of that of Lemma 4 in Section C.8. Note that $\lambda_q^2(\Psi)$ is the smallest eigenvalue of $\Psi\Psi^\intercal = \sum_{l=1}^p \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal$ and $\sum_{l=1}^p \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal - \sum_{l\in A}\Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal$ is a positive definite matrix. By the same argument for (177), we have

$$\lambda_q^2(\Psi) \geq \lambda_{\min}(\sum_{l\in A}\Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal) \gtrsim |A|\left(\lambda_{\min}(\Sigma_\Psi) - \lambda_{\max}(\Sigma_\Psi)\sqrt{q/|A|}\right). \tag{183}$$

Similarly, we have

$$\lambda_q^2(\Psi_{-j}) \gtrsim |A|\left(\lambda_{\min}(\Sigma_\Psi) - \lambda_{\max}\Sigma_\Psi)\sqrt{q/p}\right). \tag{184}$$

We establish (19) by the condition (55) on the set cardinality $|A|$.

Similarly to (179), we establish that, with probability larger than $1 - \frac{1}{t^2}$,

$$\|\Psi a\|_2^2 \lesssim t^2 q \max\{\lambda_{\max}(\Sigma_\Psi), C_1\}\|a\|_2^2.$$

Then we can establish (180), (181) and (182) by replacing $\sqrt{\lambda_{\max}(\Sigma_\Psi)}$ with $\sqrt{\max\{\lambda_{\max}(\Sigma_\Psi), C_1\}}$. Combined with (183) and (184), we establish the lemma.

## APPENDIX D: ADDITIONAL SIMULATIONS

We present here some additional simulations to the ones presented in the Section 5.1. We use the same simulation setup where we further vary certain aspects of the data generating distribution or we vary the tuning parameters of the proposed Doubly Debiased Lasso method.

*No confounding - Toeplitz and Equicorrelation covariance.* Here we explore further the scenarios where there is no confounding at all, i.e. $q = 0$, similarly as in the bottom part of Figure 7, but with different covariance structure of $X = E$. We fix $n = 300, p = 1,000$, and take the covariance matrix $\Sigma_E$ to be either a Toeplitz matrix, with $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$ for $\kappa \in [0,1)$, or we take it to be equicorrelation matrix where $(\Sigma_E)_{i,j} = \kappa \in [0,1)$ when $i \neq j$ and $1$ otherwise. In both cases, as the correlation parameter $\kappa$ approaches 1, the singular values become more spiked and the predictors become more correlated. The results can be seen in Figure A1. We see that Doubly Debiased Lasso seems to have much smaller bias $|B_\beta|$ and thus better coverage even in the case when $q = 0$, because trimming large singular values reduces the correlations between the predictors. This difference in bias and the coverage is even more clearly pronounced for the equicorrelation covariance structure, since for the Toeplitz covariance structure $\mathrm{Cor}(X_i, X_j)$ decays as $|i - j|$ gets bigger, whereas for equicorrelation case it is constant and equal to $\kappa$.

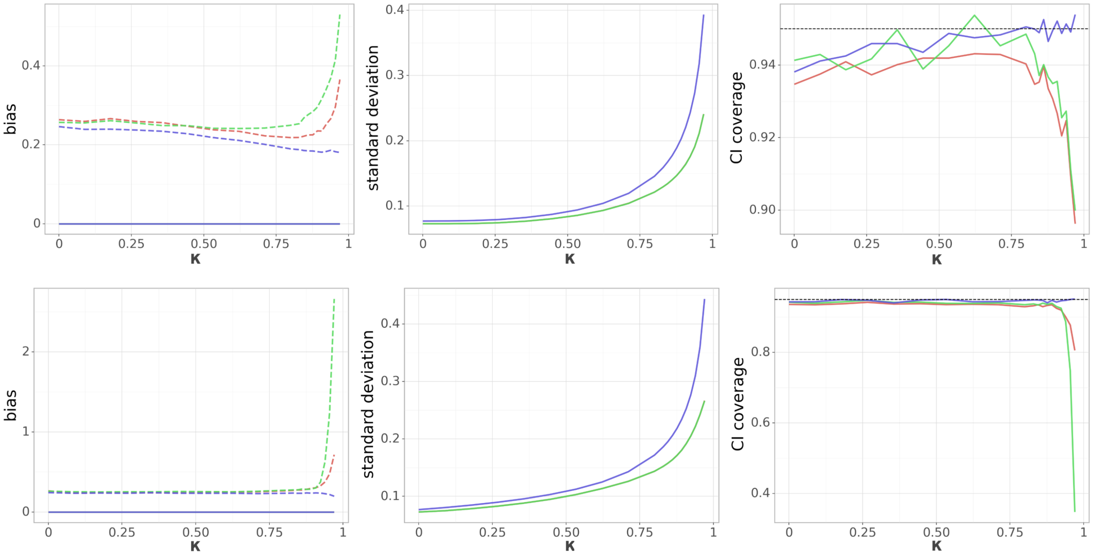

FIG A1. *(No confounding - Toeplitz and Equicorrelation covariance) Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the $95\%$ confidence interval (right) on the correlation parameter $\kappa$, while keeping $p = 1,000, n = 300, q = 0$ fixed. In the plots on the left, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively, but $B_b = 0$ since we zero confounders $q = 0$. Top row corresponds to the Toeplitz covariance structure $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$, whereas for the bottom row we have equicorrelation covariance matrix where the off-diagonal elements equal $\kappa$. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $V$.*

*Non-Gaussian distribution.* The Assumption (A3) in Section 4 requires that the noise term $\nu_{i,j} = E_{i,j} - E_{i,-j}^\intercal \gamma^E$ is is independent of $E_{i,-j}$. This condition will automatically hold if $E_{i,\cdot}$ is multivariate Gaussian or $E_{i,\cdot}$ has independent entries. We now test the robustness of Doubly Debiased Lasso method when this assumption is violated. In order to examine that, we repeat the simulation setting displayed in Figure 3, where $n = 500$ and $p$ varies from $1$ to $2,000$. We change the distribution as follows: Let $\mathbb{P}$ be some real distribution with zero mean and unit variance. The entries of the matrix of the confounders $H$ are generated i.i.d. from $\mathbb{P}$. Furthermore, the unconfounded part of the predictors $E$ is generated as $Z\Sigma_E^{1/2}$, where $Z$ is a $n \times p$ matrix with i.i.d. entries coming from the distribution $\mathbb{P}$ and $\Sigma_E$ is a Toeplitz matrix with $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$ for $\kappa = 0.7$. Finally, the noise variables $e_i$ used for generating $Y$ (see Equation 2) are also generated from $\mathbb{P}$. The results can be seen in Figure A2. We take $\mathbb{P}$ to be the following distributions: standardized chi-squared with $1$ degree of freedom, standardized t-distribution with $5$ degrees of freedom and standardized $\text{Bin}(16, 0.5)$. For comparisons of the performance, we also include $N(0, 1)$ distribution, but one needs to keep in mind that the obtained plot differs from the one in Figure 3 because of different correlation structure of $E$. We can see that there is very little change in the performance of the proposed estimator, thus showing that Doubly Debiased Lasso can be used for a wide range of models.

*Comparison to PCA adjustment.* Here we investigate how the choice of the spectral transformation can affect the performance of the Doubly Debiased Lasso estimator. We focus on the PCA adjustment which maps first $\hat{q}$ singular values to $0$, for some tuning parameter $\hat{q}$, while keeping the remaining singular values unchanged. This transformation is used frequently in

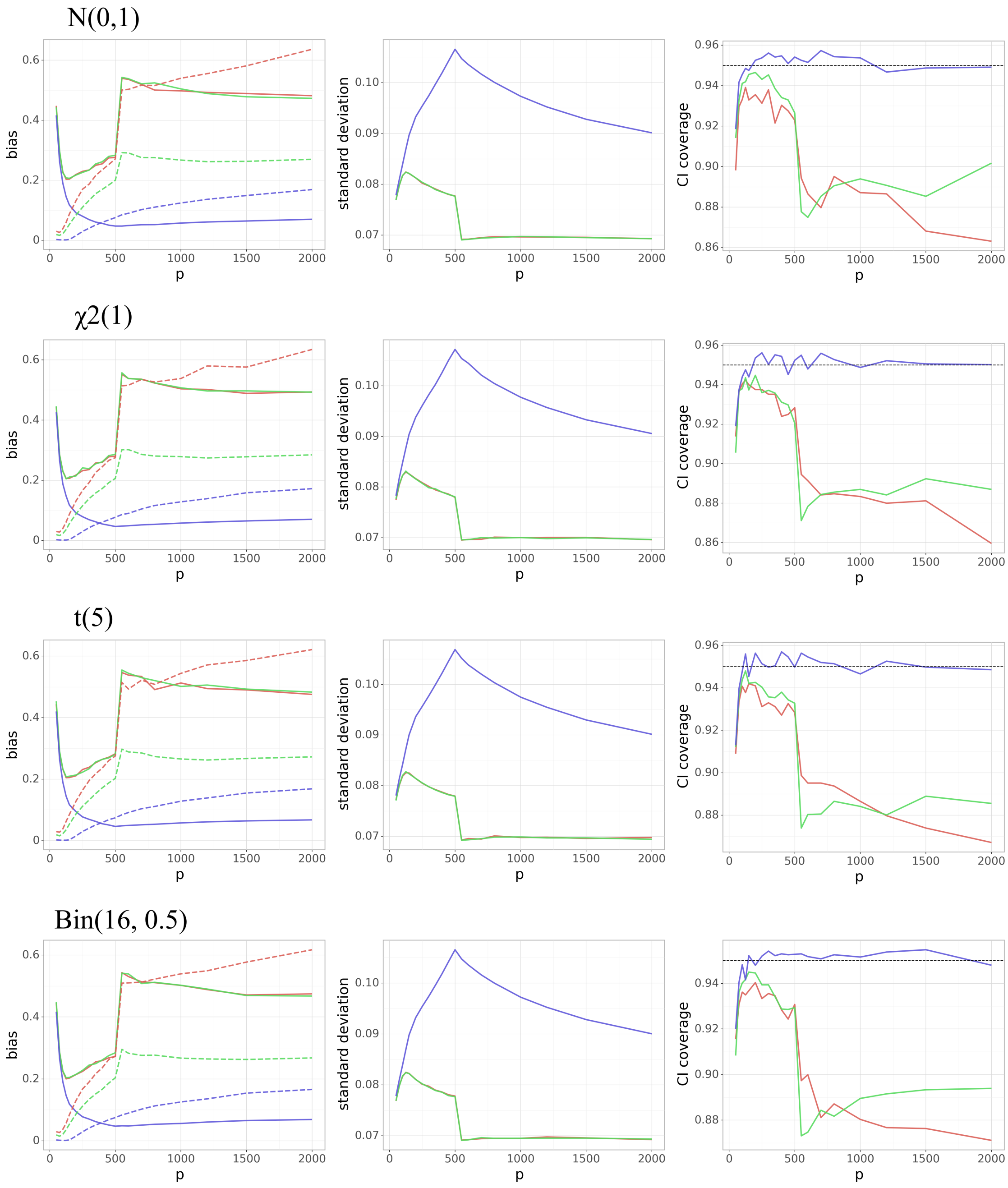

FIG A2. *(Non-Gaussian distribution) Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of predictors $p$, while keeping $n = 500, q = 3$ fixed. On the left side, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. We change the distribution of $H, E, e$ in (1) as described in the text. Each row in the plot corresponds to a different distribution $\mathbb{P}$. We set $\Sigma_E$ to have Toeplitz structure with parameter $\kappa = 0.7$. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and $V$.*

the literature because it arises by regressing out the top $\hat{q}$ principal components from every predictor.

We fix $n = 300, p = 1,000, q = 5$ and vary the parameter $\hat{q}$. We compare the estimator using the PCA adjustment for both $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ with the estimator using the Trim transform with the median rule for both $\mathcal{P}^{(j)}$ and $\mathcal{Q}$. Finally, we also consider the estimator using the Trim transform for $\mathcal{Q}$ and PCA adjustment for $\mathcal{P}^{(j)}$, in order to separate the effects of changing the spectral transformation for the initial estimator $\widehat{\beta}^{init}$ and the overall estimator construction. The results can be seen in Figure A3.

We see that the performance is very sensitive to the choice of the tuning parameter $\hat{q}$. On one hand, if $\hat{q} < q$, we do not manage to remove enough of the confounding bias $B_b$, which has as a consequence that there is certain undercoverage of the confidence intervals. On the other hand, if $\hat{q} \geq q$, the bias $B_b$ becomes very small, but the variance of our estimator increases slowly as $\hat{q}$ grows. Also, removing too many principal components when computing $\widehat{\beta}^{init}$ can remove too much signal, resulting in the higher bias $B_\beta$. Trim transform has an advantage that we do not need to estimate the number of latent confounders $q$ from the data, which might be a quite difficult task. This is done by trimming many principal components, but not removing them completely. However, this can result in a small increase of the estimator variance compared to the PCA adjustment with the optimal tuning $\hat{q} = q$.

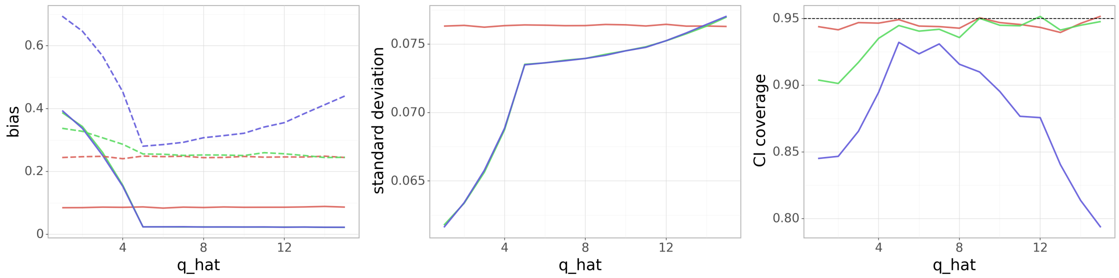

FIG A3. *(Comparison to PCA adjustment) Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the correlation parameter $\kappa$, while keeping $p = 1,000, n = 300, q = 5$ fixed. In the left plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. We vary the parameter $\hat{q}$ of the PCA adjustment, which maps the first $\hat{q}$ to zero. Red color corresponds to the Doubly Debiased Lasso using Trim transform for both $\mathcal{P}^{(j)}$ and Q, blue color represents the Doubly Debiased Lasso using PCA adjustment for both $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ and green color corresponds to the Doubly Debiased Lasso estimator using the same default $\widehat{\beta}^{init}$ with $\mathcal{Q}$ being the median Trim transform, but uses PCA adjustment for $\mathcal{P}^{(j)}$. Note that the last two methods have almost indistinguishable $V$.*

*Weak confounding.* Here, we explore how the performance of our estimator depends on the strength of the confounding, i.e. how $H$ affects $X$. In Figure 5, we have already explored how the performance of our method depends on the number of affected predictors by each confounder. Here we allow all predictors to be affected, but with decaying strength. This we achieve by generating the entries of the loading matrix $\Psi$ as $\Psi_{ij} \sim N(0, 1/\sigma_i(j)^a)$, where for each of the $q$ rows we take a random permutation $\sigma_i : \{1, \ldots, p\} \to \{1, \ldots, p\}$, and $a \geq 1$ is a tuning parameter describing the decay of the loading coefficients. The values $n = 300, p = 1,000$ and $q = 3$ are kept fixed. The results can be seen in the Figure A4. We see that when $a$ is close to 1 and the confounding is strong that our proposed estimator is much better that the standard Debiased Lasso estimator. On the other hand, when $a$ is larger, meaning that the confounding gets much weaker, the difference in performance decreases, but Doubly Debiased Lasso still has smaller bias and thus better coverage.

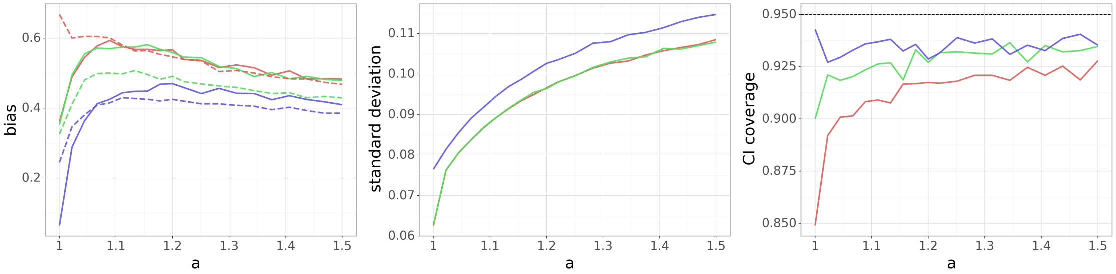

FIG A4. *(Weak confounding) Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the $95\%$ confidence interval (right) on the loadings decay parameter $a$, while keeping $p = 1,000, n = 300, q = 3$ fixed. In the left plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $V$.*